DOCUMENT RESUME

ED 104 569	PS 007 880

AUTHOR	Raizen, Senta; And Others
TITLE	Design for a National Evaluation of Social Competence in Head Start Children.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	Office of Child Development (DHEW), Washington, D.C.
REPORT NO	R-1557-HEW
PUB DATE	Nov 74
NOTE	465p.; For related documents see TM 004 561, "Toward a New Cognitive Effects Battery for Project Head Start", and PS 007998, "Appendixes," in RIESEP75
EDRS PRICE	MF-$0.76 HC-$23.48 PLUS POSTAGE
DESCRIPTORS	Cognitive Development; Evaluation Criteria; Evaluation Methods; Health; Language Development; *National Surveys; Nutrition; Perceptual Motor Learning; *Preschool Children; *Preschool Evaluation; *Research Design; Social Development; *Social Maturity; Statistical Analysis; Test Construction
IDENTIFIERS	*Project Head Start

ABSTRACT

This volume specifies the design for a national evaluation of the effects of Head Start programs on the total child, defined in terms of his social competence (in assuming the role of pupil), but is not meant to be construed as a recommendation that a national evaluation be undertaken. The first chapters contain introductory recommendations concerning the use of the evaluation design; review of many of the theoretical and methodological problems involved in determining outcome criteria and producing interpretable, socially important, and socially responsible data; discussion of background information and issues which influenced the designing of the evaluation; and an overview of the evaluation, including detailed reasons for the choices made in respect to the main elements of the evaluation design. The following chapters contain specific examinations of these areas: (1) Health and Nutrition; (2) Perceptual-Motor, Cognitive, and Language Development; (3) Social and Personal Development; and (4) Independent Variables concerning treatment, control groups, and background characteristics. The final sections of the volume include the basic evaluation design and discussions of issues of statistical analysis, test development, pilot tests of the national evaluation, and the importance of using focused (small-scale) studies as adjunct to (and perhaps instead of) a national evaluation. (SDH)

ED104569

DESIGN FOR A NATIONAL EVALUATION OF SOCIAL COMPETENCE IN HEAD START CHILDREN

PREPARED FOR THE OFFICE OF CHILD DEVELOPMENT,
DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

SENTA RAIZEN
SUE BERRYMAN BOBROW

WITH TORA KAY BIKSON
JOHN A. BUTLER
KAREN HEALD
JOAN RATTERAY

R-1557-HEW
NOVEMBER 1974

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PS 007880

Rand
SANTA MONICA, CA. 90406

This report was prepared for the Office of Child Development, Department of Health, Education, and Welfare, under Grant No. H-9766-A/H/O. Views or conclusions contained herein should not be interpreted as reflecting the official opinion of the sponsoring agency.

Cover: Stefani Relles (age five)

Published by The Rand Corporation

00003/4

PREFACE

At the request of the Office of Child Development of the Department of Health, Education, and Welfare (Grant H-9766-A/H/O), The Rand Corporation has designed a national evaluation of the social competence effects of the Head Start program. This is the final report for that study.

The Head Start program has been the subject of several major studies--among them, the Westinghouse-Ohio University evaluation and the Head Start Planned Variation evaluation. One of the purposes of a new evaluation was to examine the program's effects on the total child, defined in terms of his social competence. Another was to collect data that were more credible than earlier studies to different groups concerned with the program or its evaluation.

This report specifies the major requirements for a national evaluation that uses "social competence" as the criterion variable, respects the rights of the multiple constituents of the evaluation, and yields meaningful data. It is possible to run such an evaluation. However, we foresee many difficulties between an acceptable design and interpretable, socially important, and socially responsible data--e.g., developing measures for certain outcomes; adapting other measures to reflect the cultural variation among subjects; field problems, including community cooperation with random assignment of children to treatment and control groups and with extensive testing of children. We are not sanguine that these hurdles can be negotiated successfully. A national evaluation consumes vast resources--money, time, patience, the reputations of concepts and of people. Since it is not clear that the objectives of the evaluation can be met successfully, we believe these resources would be more fruitfully spent in a system of small, carefully designed studies of some of the major substantive and methodological problems that became apparent as we developed this evaluation design.

Although we would prefer to see the resources spent on small, careful studies, we believe that this design will be useful to the OCD, researchers interested in general issues of evaluation, policymakers who are responsible for initiating or for using the results of evaluations of educational interventions, child development experts, and groups concerned with the protection of the rights of subjects. Designing an evaluation with this criterion variable, for this subject population, and with the methodological constraints of an on-going social program forced us to confront issues of interest to all of these groups.

A number of people were involved in the design, but the report itself was written by Rand researchers and consultants. Tora Kay Bikson, Sue Berryman Bobrow, Karen Heald, and Joan Ratteray are with The Rand Corporation. John Butler is a Rand consultant and editor of the *Harvard Educational Review*. Senta Raizen, formerly with The Rand Corporation, is now with the National Institute of Education.

Senta Raizen and Sue Berryman Bobrow are responsible for Chapter 1 on background and issues of the evaluation plan; Senta Raizen, for Chapter 2 on evaluation overview. Karen Heald is responsible for Chapter 3 on health and nutrition; John Butler, for Chapter 4 on perceptual-motor, cognitive, and language domains; Tora Kay Bikson, for Chapter 5 on social and personal development; and Joan Ratteray, for Chapter 6 on the independent variables. Sue Berryman Bobrow is responsible for Chapter 7 on the design for the evaluation, Chapter 8 on statistical analysis issues, and Chapter 9 on instrument development and the pilot test of the evaluation. Sue Berryman Bobrow, Tora Kay Bikson, and John Butler collaborated to produce Chapter 10 on special studies. Senta Raizen was project director for the first year of the project; Sue Berryman Bobrow assumed this responsibility on July 1, 1974.

SUMMARY

This volume specifies the design for a national evaluation of the social competence effects of Head Start on eligible children. Data from a national evaluation are expected to be used either to evaluate or to shed light on these four questions:

1. What are the social competence effects of Head Start for members of the eligible population who receive the treatment, relative to members of that population who do not?
2. What are the social competence effects of Head Start for eligible children from different cultural groups who receive the treatment, compared to eligible children from those same groups who do not?
3. What are the social competence effects of Head Start for eligible children within each cultural group who receive the treatment and who differ in entry characteristics, as indicated by pretests and other background characteristics?
4. Are there any indications that variations in treatment produce variations in social competence outcomes for children who receive the treatment?

Our major recommendations are given first. Then we discuss, in order, the technical design of the evaluation, the independent and dependent variables, and the statistical analysis.

MAJOR RECOMMENDATIONS

This volume sets forth what we believe to be the requirements for acquiring interpretable, socially meaningful data on the social competence effects of Head Start for eligible children. However, a *plan* for a national evaluation should not be interpreted as a recommendation that such an evaluation be conducted. We foresee many difficulties between an acceptable design and the acquisition of interpretable and meaningful data--e.g., developing measures that are comparable across subcultural groups; obtaining community support for random assignment of children to treatment and control groups. We doubt that all of these hurdles can be negotiated successfully, and the failure to do so will seriously diminish the quality of the results. Our strong preference is to invest the substantial resources that would be consumed by a national evaluation in a system of small, careful studies addressed to the substantive and methodological issues that surfaced in the course of designing the national evaluation. Chapter 10 describes several special studies that can be viewed either as adjuncts to a national evaluation or as elements in such a system of small studies. Small, rigorous studies can contribute to evaluating *national* effects by elucidating: (1) the relationships among the variables that affect children's outcomes; and (2) the conditions under which we expect these relationships to shift.

If, as we recommend, a national evaluation is not undertaken, we believe that this design will nonetheless be useful to the OCD, to researchers interested in general issues of evaluation, to policymakers who are responsible for initiating or for using the results of evaluations of educational interventions, to child development experts, and to groups concerned with the protection of the rights of parties to an evaluation. Designing an evaluation with a criterion variable that is difficult to define and measure, for a multi-ethnic and young population, and with the methodological constraints of an on-going social program forced us to confront issues of interest to all of these groups.

In the course of designing the evaluation, we encountered three issues that cut across the total evaluation plan: definition of the criterion variable, "social competence"; the feasibility and appropriateness of a longitudinal assessment of social competence; and definition of the constituents of a national evaluation and their rights. If the evaluation plan is implemented, we recommend that these issues be resolved as follows: "social competence" should be defined as competence in the role of pupil and as those variables that can be expected to affect competence in that role; a longitudinal evaluation is unlikely to be feasible, implies an inappropriate expectation for Head Start effects, and should not be done; and a national evaluation of a social program should be defined as a multiple-constituency process and structured to protect the rights of those constituents.

We encountered major problems in defining the criterion variable, "social competence," for these reasons: It does not have precise meaning within the society as its meaning varies among and within subcultural groups; it does not have precise meaning within child development theory; and, as an alternative basis for definition, the necessary and sufficient conditions for "social competence" are not clearly understood within child development theory. One solution to the definitional problem is to argue that a monolithic definition of social competence is not fruitful: A child is perceived as socially competent or incompetent in the context of specific roles and value judgments. From the point of view of designing a national evaluation, the question is: What roles can we expect Head Start to affect--i.e., what roles are available in Head Start, and in which ones might Head Start-eligible children be expected to be less competent than other children? Head Start does not offer the child roles in the family or community, and there is no reason to think these children lack competence in these roles. Head Start does make available a foundation for the socially important role of pupil, and we can expect eligible children to be less prepared for that role than, say, middle-class children. Thus, the criterion of Head Start success becomes Head Start effects on children's competence in assuming the role of pupil.

The other solution to the definitional problem is to equate social competence with dependent variables that seem to produce effectiveness in the role of pupil. This definition leads us to groups of health and nutrition variables; perceptual-motor, cognitive, and linguistic variables; and social and personal variables.

Our next recommendation is that if a national evaluation is conducted, it should not be a longitudinal evaluation. The Office of Child Development was originally interested in this possibility. However, we advise against it for several reasons: serious theoretical problems with linking outcomes produced by Head Start and outcomes at a later stage in the child's development; cost of following up a nationally representative sample for three or four years; diminished cell sizes when the data are controlled at least for the child's school, and possibly for his classroom. We also argue that Head Start is appropriately held accountable for only immediate, not long-term, effects. We assume that the development of a child is a somewhat Markovian process--i.e., outcomes at time_n are primarily determined by immediately preceding outcomes at time_{n-1}, not so much by those at time_{n-2} or time_{n-3}. For example, if a child has reading problems in third grade, it is probably more appropriate to hold the second grade responsible than Head Start.

Our third recommendation is that an evaluation process be structured as a multiple-constituency process, with explicit recognition of the rights and obligations of each constituent in that process. We see the constituents as Head Start-eligible children, the parents of those children, the cultural groups to which the children belong, the personnel responsible for the program at local levels (e.g., Head Start teachers, center directors, program directors), the federal agencies responsible for the program (e.g., Office of Child Development, Assistant Secretary for Planning and Evaluation), the scientific community, the scientists who conduct the evaluation, and the taxpayer as represented by the Congress. We specify major rights for these groups in Chapter 1--e.g., right of all constituents, including Head Start parents and program personnel, to define alternative criteria for Head Start success.

If the design is implemented, a last major recommendation is to set aside 18 months to prepare for the full-scale evaluation. (Six months of this period overlap the first part of the full-scale evaluation.) We see this period as essential for developing and adapting measures and for conducting a pilot test of all procedures and measures of the evaluation. In Chapter 9 we specify criteria for accepting or rejecting newly developed measures for the pilot test. We also specify the mechanics of the pilot test and criteria for certain choices that have to be made in the course of it. We urge that the pilot test experience be used to make several critical go no-go decisions--e.g., to screen out measures whose reliability or validity is problematic; to choose optimal sample sizes for numbers of children per site and number of sites; and, *most important*, to review the feasibility and substantive promise of the full-scale evaluation.

TECHNICAL DESIGN

The Head Start program places important constraints on a technical design. As indicated in Chapter 7, we expect considerable variation within each treatment level--i.e., within Head Start and within all alternatives to Head Start. Even if the variations can be distinguished, they cannot be randomly assigned to sites. Thus, they are almost certainly confounded with sites.

A second constraint is that children cannot be randomly assigned across sites. We can then expect confounding of site effects and child background characteristics.

Since the design cannot disentangle variations within treatment and control conditions from site effects and cannot separate child characteristics from site effects, it is not structured to yield *a priori* estimates for questions 3 and 4 (see page v, above). These questions can still be evaluated, but in *ex post facto* ways. The design is structured to yield unbiased *a priori* estimates for questions 1 and 2 by requiring that (1) children are randomly assigned to control and treatment conditions within site, (2) effects are estimated separately for individual sites, and (3) effects are aggregated across sites.

Thus, for the technical design we recommend a two-stage cluster sample of Head Start centers and classrooms within centers. We also recommend random assignment of eligible volunteers for the Head Start program within each site to treatment and control groups and to classrooms within the treatment group. Although a random assignment design is preferred from a set of several alternatives, design based on growth curves, called a "value-added" design, is considered promising.

We expect variations in Head Start and in the alternatives to Head Start (control) across sites and variations among children across sites. Therefore, we wanted to use a design that stratified treatment and control conditions and children. Since we have no theoretical basis for differentiating variations within Head Start and its alternatives (control), we were unable to stratify the design beyond Head Start and not-Head Start conditions.

We advise stratifying the population of eligible children by selected demographic properties of Head Start catchment areas: ethnic origin of children served by the center, a version of the urban-rural dimension, and region.

INDEPENDENT AND DEPENDENT VARIABLES

The independent and dependent variables selected for the design are summarized in the tables in Chapter 2. The independent variables (Chapter 6) involve "trial" variables for differentiating variations within the Head Start program and within the control condition. Since the theory of classroom process is not well developed, the variables for identifying meaningful variations in Head Start classrooms are only tentatively recommended. We also advise measuring a number of variables about child, family, teacher, Head Start center, and community as sources of explanations for variations in children's outcomes within a site or across sites.

The dependent variables fall into three groups--health and nutrition (Chapter 3); perceptual-motor, cognitive, and linguistic (Chapter 4); and social-psychological (Chapter 5). The ages of Head Start children--3 to 5 years--are among the healthiest of the life span. However, these particular children are susceptible to problems that Head Start can affect: infectious diseases for which there are immunizations; nutritional deficiencies, especially iron; undetected auditory and visual problems; and untended dental decay.

At least some of the effect of any good Head Start program is reflected in what the *OCD-Head Start Policy Manual* (1973, p. 7) terms "the enhancement of the child's mental processes and skills with particular attention to conceptual and verbal skills." The perceptual-motor, cognitive, and language skills considered in Chapter 4 contribute to our impression of the child's social competence, constituting what for many are the most important dimensions of program-related growth. We selected outcomes in these three domains according to the following criteria: Is the outcome teachable: Is it one that developmentally intact children of Head Start age can be expected to learn? Is it an outcome that Head Start can be expected to teach? Is it one that in the absence of Head Start, eligible children can be expected to learn less well or not to learn? In other words, there is an emphasis on the assessment of "leverageable" and behavioral areas of program effect, such as visually guided fine motor skills, or comprehension and recall of verbal communication. Dimensions of mental capacity normally explored in IQ and other stable trait measurement are given low priority.

In selecting measures of perceptual-motor, cognitive, and language outcomes, we use three criteria: The measure is a valid indicator of the dependent variable, it has high reliability under field conditions, and the child's performance required by the specific measure has *in itself* social meaning for constituents of the evaluation. On the basis of these criteria we recommend a subset of the CIRCUS test battery, developed by the Educational Testing Service. Chapter 4 reviews the conceptual and technical characteristics of these tests. We also recommend that one measure be designed and pretested for use in the evaluation: a two-person communication game, to assess referential language use in the Head Start classroom.

One measure, the Ravens matrices, does not explore "leverageable" aspects of learning and skill acquisition. This measure is included as a maturational indicator or developmental baseline marker of at least one stable dimension of intellectual ability. We recommend that this measure be used with other background variables as a covariate or other control in analyses of program effect. In itself, it is *not* expected to be used as a gain measure.

In the selection of the social and personal variables (Chapter 5), we assume that the child who has developmentally intact skills is asked to perform in three kinds of statuses: specific statuses to which any child might aspire, such as the role of successful student; ascribed statuses, such as age, sex, and social class statuses; and individual statuses that the child can invent or negotiate for himself within the constraints of the ascribed statuses. It is also assumed that the styles of behavior learned as appropriate by the lower socioeconomic class child and that are successful within certain social environments are not necessarily associated with success in the role of pupil. One of the purposes of Head Start is to help the child elaborate individual statuses that allow him to integrate roles associated with his ascribed status with a successful enactment of the role of pupil.

The variables are conceived of as situationally learned responses, not as intra-individual traits. Since there is no generally accepted model of social and personal competence, these outcomes could not be derived from established theory. The most important outcomes are the child's behaviors toward other people who are significant in the school environment (peers and teachers) and the "feedback" behaviors (expectations and evaluations) of these same persons toward the child. A second set of outomes involves those responses of children that affect academic performance (e.g., test-taking behavior, curiosity) and styles associated with more general successful social behavior (e.g., *range* of responses, as indicated by the child's ability to generate multiple solutions to a problem). Attitudes are given a low priority. They are difficult to measure for this age group; paper and pencil tests are the standard measurement technique in this area, and these are useless for young children. It is also difficult to establish links between attitudes and behavior. Only attitudes toward school and attitudes toward self are recommended for measurement.

We recommend that the health and nutrition variables and perceptual-motor, cognitive, and language variables be measured for the treatment and control groups at the close of the Head Start year. The effects of Head Start on the child's ability to handle the socioemotional aspects of the role of pupil, however, are most visible in the school context if they are measured before they become confounded with the school experience itself. Thus, we recommend the evaluation of these variables early in the first public school year and in the public school context.

STATISTICAL ANALYSIS

Statistical analysis issues are discussed in Chapter 8. The chapter does not recommend specific statistical tests; the group analyzing the data will be as familiar as we are with particular statistical descriptors and tests. We recommend instead ways of *thinking about* statistically defensible and socially meaningful analysis of evaluation data and its combination to the policy world. For example, the distinction between confirmatory and exploratory data analysis is discussed. We recommend confirmatory data analysis for evaluating questions 1 and 2, both of which have *a priori* properties, and exploratory data analysis for questions 3 and 4, both of which have *a posteriori* properties. Since national evaluations are conducted to test specific assumptions about program effects, we recommend that substantive questions be evaluated by significance tests as well as by confidence intervals.

In dealing with the form of the hypothesis we recommend a two-tailed form of the "null" and alternative hypotheses and suggest Neyman's criterion as the basis for selecting the specific statement of the null. In discussions of analysis models for a random assignment design, we strongly recommend that each site (Head Start catchment area) be analyzed as a separate experiment. Effects of Head Start across sites are then estimated by aggregating site-specific effects.

CONCLUSION

If these design recommendations are implemented, we think that interpretable, socially meaningful, and socially responsible data can be obtained. However, we reiterate our belief that some recommendations will be difficult to implement at all, let alone sufficiently. We also reiterate our preference for using the resources that would be consumed by a national evaluation in a system of small, carefully designed and executed studies.

ACKNOWLEDGMENTS

This report could not have been written without help from many people. To begin with, thanks are due to the authors of issue papers in four areas of child development: Roslyn Alfin-Slater and D. B. Jelliffe for health and nutrition, Herbert Pick for motor/perceptual development, Helen Featherstone for language development, and John A. Butler for cognitive development. We have also profited greatly from consultation with other persons and from panel discussions and recommendations in these and a number of related areas. Panel participants and individual consultants are listed in Appendix A. The contributions made to this study by panels of Black and Spanish-surnamed professionals are gratefully acknowledged; their contributions are abstracted in Appendix B.

Helen Bee Douglas and Theodore Donaldson served as the internal Rand reviewers of the report. Their comments were very helpful in locating problems in the draft, and we appreciate the care they took in the process.

Gratitude is also owed to R. Victoria Robinson for her assistance with the panels and many other aspects of the project and to Betty Ann Burness and Luetta Stevens for their substantial secretarial help.

CONTENTS

Chapter

1. Background and Issues 1
2. Evaluation Overview 31
3. Health and Nutrition 69
4. Perceptual-Motor, Cognitive, and Language Outcomes and Measures 93
5. Social and Personal Development 151
6. Independent Variables 235
7. Basic Evaluation Design 264
8. Issues of Statistical Analysis 317
9. Test Development and Pilot Test of the National Evaluation 339
10. Focused Studies 386

Appendix

A. Panel Participants and Consultants in the Rand Head Start Project 411
B. Abstract of Recommendations of Black and Spanish-Surnamed Professionals' Panel 421

BIBLIOGRAPHY 427

Chapter 1

BACKGROUND AND ISSUES

HISTORY 2

LESSONS TO BE DRAWN 9
  Lack of Cumulation 9
  Inadequacy of Instrumentation 9
  Inferential Problems 13
  Field Problems 14

ISSUES IN THE DESIGN OF A NEW NATIONAL EVALUATION 14
  Social Competence 15
  Payoffs from a Longitudinal Study 19
  Evaluation as Multiple-Constituency Process 23
  Payoffs from a National Evaluation 25

Chapter 1

BACKGROUND AND ISSUES

HISTORY

In July 1973, Rand was awarded a grant from the Office of Child Development (OCD) to design a comprehensive evaluation of Head Start. The proposed evaluation was to have the following objectives:

- To yield information on how well the Head Start program is achieving the goal of improving the social competence of disadvantaged children.
- To yield information on how the Head Start program might be improved to accomplish its child development goals more effectively.

The decision by OCD to initiate a new evaluation appeared to be based on the following considerations:

- Head Start has evolved considerably since the last attempt to assess its national effects in the Westinghouse-Ohio Study (Cicarelli et al., 1969). Its current goals have been outlined and minimum performance standards established in the *OCD-Head Start Policy Manual* (January 1973). Any conclusions about the program derived from past studies are therefore outdated.
- Previous evaluations took too narrow a view of Head Start's goals in child development, focusing largely on cognitive gains. This criticism was prominently articulated by Zigler (1973), the former director of OCD, and has provided major impetus for a new evaluation that would consider a much broader range of outcomes, particularly those important to the disadvantaged child's social competence.

- Conclusions from earlier studies were open to attack--and, in fact, heavily criticized--because of the weak inferential base provided. Flaws included the use of inappropriate measures, the subversion of initial evaluation goals by field problems, faulty assumptions with respect to program characteristics (particularly in the Planned Variation studies), study designs allowing alternative explanations of any effects found by the study, and inadequate analytical tools. Hence, despite previous efforts, little definitive information could be given to the insistent questions on what the investment of over $400 million a year in Head Start was accomplishing for participating children.

These considerations were reflected in the parameters originally established by OCD for the proposed new evaluation. *First*, the evaluation was to focus on gains made by disadvantaged children in social competence as a result of participating in Head Start, where social competence was defined by OCD to mean (*OCD-Head Start Policy Manual*, p. 6):

> the child's everyday effectiveness in dealing with his environment and later responsibilities in school and life. Social competence takes into account the interrelatedness of cognitive and intellectual development, physical and mental health, nutritional needs, and other factors that enable a child to function optimally.

The emphasis was to be on child outcomes; other Head Start goals relating to improvement of conditions for the participating families or communities and to better functioning of other service agencies were to be addressed outside of this particular evaluation effort. *Second*, the evaluation was to include a representative sampling of the populations participating in Head Start. *Third*, it was to demonstrate both short-term and long-term effects of Head Start; therefore a five-year longitudinal study was proposed during which two succeeding waves of Head Start children would be followed up through third grade. *Fourth*, in order to obtain results as rapidly as possible, data collection on the first wave was to begin in fall 1974, on the basis of the design work by Rand. On the strength of Rand's interim recommendations made in December 1973, OCD agreed to reschedule the full-scale evaluation to fall 1975 so as to allow *a preparatory year for needed test development, community coordination, and piloting of the recommended test battery and design.*

Since the proposed new evaluation was conceived partly in response to disappointing results from previous studies, it is important to understand their development and shortcomings. Head Start has, in fact, been exposed to more critical scrutiny than almost any other single federal program. Since its inception within the Office of Economic Opportunity (OEO) in 1965, the program has been subject to a continuing variety of examinations of its functions and its worth. As one considers the resulting studies and their evolution, one is struck by the attempt in each succeeding effort to do a better job, either by constructing a revised set of assumptions or by avoiding previous mistakes. Any new evaluation must be similarly informed by the experience accumulated through this history.

During the first phase, from 1965 to 1969, with the cooperation of the Bureau of the Census, descriptive surveys of a nationally representative sample of Head Start centers were conducted to measure the degree of compliance with Head Start guidelines. Overall compliance was found to be high. Also, a profusion of individual research studies was funded by OEO in an effort to evaluate the effects of the Head Start experience on the children served by the program. However, these studies were conceived in isolation, without fitting into any comprehensive evaluation framework, and were conducted on a small scale, at one or two centers, with fewer than 100 children and few measures. Although this diversity in the research led to difficulty in drawing conclusions on Head Start effects, some general effects could be summarized. Williams and Evans cite the following results as emerging from the studies by 1967:[1]

---

[1] Williams and Evans, "The Politics of Evaluation: The Case of Head Start," in Rossi and Williams (1972, p. 253).

> Across a wide range of these projects it was found that, in general, participants who had been given various cognitive and affective tests at the beginning of the Head Start program showed gains when tested again at the end of the program. However, virtually all the follow-up studies found that any differences which had been observed between the Head Start and control groups immediately after the end of Head Start were largely gone by the end of the first year of school.

The Westinghouse-Ohio Study (Cicarelli et al., 1969), sponsored by a separate evaluation office within OEO, is the only attempt to assess the national effect of Head Start on a sample representative of the overall enrollment in the program. This evaluation placed heavy emphasis on the intellectual and psychological development of participating children; the medical, nutritional, and community goals of Head Start were not addressed. The basic question the study attempted to answer was whether changes in cognitive performance were higher for children who had participated in Head Start programs than in control groups without the Head Start experience. The study used an *ex post facto* design: Children in 104 sites who had attended Head Start and were in the first, second, and third grades in the period 1968-1969 were compared with a matched sample of children who had not attended Head Start. Because of all the methodological problems inherent in *ex post facto* designs (Campbell and Stanley, 1963; Campbell and Erlebacher, 1970) the report caused considerable controversy in the scientific community (Smith and Bissell, 1970; Madow, 1969). Of greater interest to decisionmakers were the findings--disappointing with respect to original expectations--that although the (full-year) Head Start children showed somewhat greater gains initially than the control groups on the three cognitive measures used, this advantage was not maintained. No increase in cognitive gains was documentable for summer-only participants; nor did children from either summer or full-year programs score significantly higher than the control groups on the three affective measures used. Despite the controversy engendered by the report, its findings were consonant with those of the earlier research studies. (It has, in fact, been conjectured that the methodological attacks on the study might have been considerably milder if the findings had been more favorable to Head Start.) The *reasons* for the findings remained unresolved--i.e., whether critical independent and dependent variables were left out or poorly measured, whether the results were uninterpretable because of serious flaws in the design, or whether the expectations for Head Start were inappropriate. One reasonable conjecture for the failure to find greater effects is that the study dealt with programs only at the highest levels of aggregation. More effective programs were combined with less effective ones; scores were averaged across children with only a few analyses taking account of gross ethno-demographic distinctions; the part of the sample representing full-year participation was small. Thus, it was perhaps not surprising to find that overall mean gains were not very great.

A concurrent set of national evaluations sponsored by Head Start itself focused on the interactions between program and child, with the emphasis as much on program and teacher characteristics and classroom process as on measures of performance. One major evaluation stems from the efforts of a network of 14 university-based Head Start Evaluation and Research (E&R) Centers established in 1966. Research Triangle Institute (December 1972) analyzed data collected by E&R Centers in two national samples of Head Start classes (1967-1968; 1968-1969). Data included extensive descriptions of the classroom; information on child, family, and program characteristics; and personal, social, and cognitive measures of development. In general, gains on all cognitive measures were noted; cognitive measures were more susceptible to program variation than noncognitive measures; highly structured, focused, well-implemented programs produced higher gains. However, confidence in the conclusions of this study is limited by methodological considerations. First, because of the interest in within-Head Start program variation, no comparison was made with control groups of non-Head Start children. Second, the study design did not include random selection of programs and random assignment of children to programs; therefore, the programs are not necessarily representative of the total Head Start program, and characteristics of children are confounded with characteristics of programs. Lack of uniformity in testing times and testing intervals further weakens the internal validity of the design. A second analysis based on the same E&R data was conducted by System Development Corporation (May and August 1972), treating the three years of data collection (1966-1969) as three distinct studies. Adding to the list of limitations, SDC points to changes over the years in measures used, in sampling procedures followed, and in degree of structure imposed on Head Start intervention strategies (natural versus planned variations). Missing data compound interpretation problems.

One consequence of the Westinghouse-Ohio report and the E&R analyses was to focus more sharply on what kinds of programs would produce the greatest gains. Planned from 1967-68, the Follow-Through program had been conceived as an effort to provide a coherent program beyond Head Start in order to sustain Head Start gains. The two goals merged when well-defined curriculum models (Planned Variations) were applied both to Head Start and to Follow-Through programs. In 1970, Stanford Research Institute was contracted to assess the cumulative influence of such a coherent program and to compare the short- and long-term effectiveness of the various models. Observing three waves of children in 16 target communities, SRI compared four groups of children: those in regular Head Start programs or Head Start Planned Variation (HSPV) programs who went on to Follow-Through or non-Follow-Through primary school programs. As in previous national evaluations, SRI's study was plagued by the lack of randomness in assignment of HSPV programs to communities and classes, nonequivalent comparison groups, changes in test battery, and gaps in the data. In addition, while some models stressed socioemotional development, there were no adequate measures for this domain.

After the first year of data collection, control of the data analyses was transferred to the Huron Institute, though SRI continued in charge of the fieldwork. While a few socioemotional measures were added for the second and third wave, Huron's evaluation of the quality of these instruments (Walker et al., 1973) points to their poor psychometric properties and lack of validity. In fact, the preliminary report on the second wave of children (Smith, 1973) presents data on only four measures of cognitive growth and one measure of motor control because of the "sparsity and limitations of measures used in this study." In the analysis of third-wave data (Weisberg, 1973), data for only five of 12 outcome measures could be used, "because...there were crippling limitations on the usefulness and appropriateness of the other measures as evaluative instruments.... Not surprisingly, the tests which are suitable [for analysis] all measure skills in the cognitive domain." The findings of this complex study tend to confirm earlier findings of initial gains in cognitive performance for children in Head Start. Programs structured toward attainment of specified cognitive skills show greater effectiveness with respect to those skills. Conclusions on how well these gains are maintained beyond Head Start must await analyses of the Follow-Through data. Two current concerns remain unaddressed: By default, the emphasis once again turned out to be on traditional achievement measures; by design, the effort was not concerned with generalizability of findings over the total Head Start population.

To provide better information on the factors influencing child development, the Educational Testing Service (ETS) was commissioned to do a follow-up study of children from four to eight years of age. The aim was to study the cognitive, personal, and social development of disadvantaged children and to identify those components of early education that are associated with developmental gains. The original sample included both Head Start and non-Head Start children in four sites (now reduced to three); representativeness with respect to the overall Head Start population was not a criterion for sample selection. Although the study is to provide evaluative information as well, the main emphasis is an in-depth study of child characteristics and educational outcomes often associated with the effects of poverty so that more informed educational decisions can be made. Because of the wealth of data collected on both dependent and independent variables, many different kinds of analyses should eventually be possible. Detailed reports (ETS, various years, 1968-1973) are already available on the extensive test battery that contains tests especially constructed for this study, and some preliminary findings. Children in low-income families are not a homogeneous group, nor are their families. But children from poverty backgrounds often do appear to develop more slowly than their middle-class counterparts. The wide range of individual differences suggests a need for an increase in individualized instruction (*ETS Developments*, spring 1974).

LESSONS TO BE DRAWN

Even this brief outline of Head Start evaluations documents the considerable investment of intellectual and fiscal resources that have gone into assessing the effects of the program on children and on improving the program itself. The purposes of these studies have been diverse, and the resulting findings do not yield adequate information on the overall effects of the current program. Nevertheless, the wealth of experience acquired through this history provides invaluable guidance for any new study to be undertaken. What are the lessons to be drawn, and what is their implication for the proposed evaluation?

Lack of Cumulation

While each succeeding wave of studies tried to compensate for some of the deficiencies of earlier efforts, it has not been possible to build a growing knowledge base from the findings. To date, there has been no analysis to see if data on some measures could be pooled given evidence that treatments and populations tested were equivalent. Unfortunately, little information is available on treatment variables, particularly in some of the earlier studies. Secondary analyses with respect to different outcomes for different children have also been made difficult by a lack of comparability from study to study. This may be because of legitimately changing conceptual frameworks, unwarranted assumptions that had to be discarded, dissimilar samples, different test batteries (even within the same effort such as the E&R studies or the successive studies of HSPV). Whatever the reasons, there has been little systematic accumulation of information needed for understanding the overall effects of Head Start on participating children and for program improvement.

Inadequacy of Instrumentation

This problem has already been alluded to in the history of attempts to measure child outcomes in the domains of personal and social development: Time and time again, the data resulting from the use of existing measures of traits generally considered important (e.g., self-concept, locus of control, creativity, motivation, anxiety) have either had to be discarded in the analytical phase or have yielded equivocal and contradictory results (see Chapter 5). Similar experiences characterize the use of tests specially developed for individual studies, although the ETS data, when fully analyzed, may perhaps yield less disappointing results. With respect to outcomes in these affective areas, it is not easy to resolve the assessment dilemma. The problem resides not so much in construction of valid and reliable tests, even though the difficulty of creating such tests can hardly be exaggerated, particularly when the instruments are to be used on large samples by unsophisticated testers; the more fundamental problems are the lack of adequate theory with respect to social and personal development of young children and the values attached to various child characteristics. Valued personality attributes vary among cultures and among population groups within a culture, depending on each group's patterns of coping with the surrounding environment. Much of the perception as to what degree, for example, of aggressiveness or cooperation is optimal for a child of a given background depends largely on the situation and on the observer's own values. Hence, the notion of objective measurement and fixed attribution of value to a gain score for a particular outcome has meaning only in the specified context in which the child is observed and with respect to his own culture's values. Thus, the problem of not knowing (because of lack of adequate theory) which attributes to assess as most important to the child's development is confounded by the need to look differently at individual children in different situations. Even when choices of what to measure are made, many of the tests that can be administered to young children turn out to be either unreliable or inappropriate proxies for the outcome of interest.

With respect to the cognitive domain, the measurement problem is of a different nature. Many instruments are available to assess gains in cognitive development and have been used throughout the Head Start evaluation efforts. The standard instruments have adequate reliability--at least for samples with little representation of poor and minority children--and have acquired face validity. The cognitive batteries used in the early studies included *IQ tests* and, for follow-up in the elementary grades, *standard achievement tests*. Later on, attempts were made to extract components of standard IQ tests and use them as proxies for outcomes considered to be precursors to later cognitive achievement. For this purpose, adaptations of *diagnostic tests* have frequently been used. None of these three types of tests appears particularly appropriate for the following reasons:

- *IQ tests*. As Butler (1974) points out, these measures have acquired great political validity but their use can be questioned on grounds of both validity and appropriateness for the Head Start population. With respect to validity, such instruments as the Stanford-Binet were designed to measure the presumed stable trait of general intelligence, not program-related gains in performance. Hence, the meaning of *gains* on the test for young children is unclear, as contrasted to the validity that single measurements with the test have in predicting school achievement, especially when the test is administered at later ages with concomitant increases in reliability. Further, such tests encourage comparisons to norms derived from the majority population, which is not representative of the children participating in Head Start. Hence, the implicit outcome being measured when IQ tests are used and scores are compared to norms is: How effective is Head Start in changing poverty children to be like middle-class children with respect to the hypothesized "g," the presumed stable latent trait of intelligence?
- *Standard achievement tests*. Such tests have generally been used to document whether Head Start has had any lasting effect on the child's school readiness and academic performance in the lower elementary grades. The tests have validity with respect to teacher and school expectations of cognitive achievement. But

again, they can be questioned on grounds of appropriateness for the Head Start populations, since they may be much poorer predictors of the academic performance of disadvantaged children than of the performance of middle-class children. One reason is that items often deal with material more familiar to middle-class children, so that opportunity to learn (amount of exposure) is confounded with ability to achieve--a criticism that applies to IQ tests as well. Second, they are generally constructed so as to measure differences between persons, rather than growth over time of an individual. As Carver (1974) points out, most tests designed according to good psychometric principles of maximizing for individual differences will be poor for evaluating educational growth.

- *Diagnostic tests*. The third category of instruments often used has included adaptations from tests originally developed to diagnose sensory or neurological impairment--for example, the Bender-Gestalt test. Use of such tests appears to make the implicit assumption that the children eligible for Head Start are deficient in those attributes deemed important for later cognitive performance. This may lead to unwarranted and injurious labeling of competent children, without much hope of remediation for those who might benefit. Further, many of these tests require highly skilled personnel with psychological training for administration and scoring.

A special difficulty with almost all cognitive measures is that performance even for non-language skills is often as much a function of the child's ability to understand the language in which instructions are given as of his ability in the specific skill. Thus, the score does not represent true score because of cultural rather than cognitive factors, and cultural bias is introduced. This is patently true for children who do not speak English, but it also applies when the tester and the child speak the same language or dialect. Clearly, one criterion for test selection must be some built-in safety devices to assure that the child actually understands what he is being requested to do.

Inferential Problems

As was pointed out above, there has been only one study devoted to assessing the overall influence of Head Start. This study has been rightly criticized because its methodology was such that the pre-treatment comparability of Head Start and control groups remains in question. It can be--and has been--argued that Head Start and control children differed at pre-treatment on the dependent variables measured at post-treatment. If the control children started "higher" on the dependent variables than the Head Start children, the effects of Head Start are underestimated. If they started lower, effects are overestimated. Even when control groups are pretested, as they were in the Westinghouse-Ohio Study, the use of matching or statistical techniques to make treatment and control groups comparable with respect to entering profiles can be questioned on several grounds, one being that important differences leading to post-treatment differences were not measured. Whatever differences occur between children who volunteer for Head Start and control children who do not may be attributed to unmeasured variables associated with the voluntary factor, not with the treatment. At present, the most powerful technique for strengthening the inference chain that permits linkage of the Head Start experience to developmental outcomes is random assignment to treatment. A recent bibliography of randomized field experiments for program planning and evaluation (Boruch, 1974) lists nearly 40 entries in the area of training and education programs, ranging from small samples in educational experiments to large samples drawn for evaluation purposes. The author claims that only a few of the randomized experiments listed failed to be implemented completely, and the documentation of outright failures, though they undoubtedly occurred, is somewhat scarce. In the author's view, "the contents of this bibliography serve as empirical evidence for the contention that randomized experimental tests are feasible in a variety of program settings" (p. 2).

Field Problems

The yield of an evaluation is more often diminished by inadequate control over fieldwork than by conceptual difficulties, lack of appropriate instrumentation, and inadequacy of the design. Problems can occur at any number of levels: inadequate preparatory activities for parents and Head Start staff in the evaluation sites; inadequate training and supervision of testers, including standardization of familiarization activities, test administration, and scoring; slippage in testing schedules, etc. Past experience indicates that the more complex the test battery to be administered, the more pronounced the field problems. Further, the integrity of the field operation is often undermined by shifting priorities that appear to demand addition of new sites or tests for political reasons, and by the inadequate handling of those originally included because of pressures for results. Cost-effectiveness considerations dictate that sufficient funds must be invested to control the quality of field operations, even at the cost of reducing the test battery and the number of subjects, and therefore the representativeness of the overall sample.

ISSUES IN THE DESIGN OF A NEW NATIONAL EVALUATION

Earlier evaluations of the Head Start program--e.g., Westinghouse-Ohio, Head Start Planned Variation--illuminated the problems of inappropriate or unreliable measurement, weak design, and field problems. In the course of designing the new evaluation, we tried to deal not only with these difficulties but also with problems unique to this evaluation, as the result of OCD requirements or exacerbated by the history of evaluation in general and Head Start evaluation in particular. These problems arose as the result of trying to:

- o Use "social competence" as the criterion of success for the program;
- o Evaluate *longitudinal* effect;
- o Redefine a national evaluation of a major social program as a *multiple-constituency process*;

- Use the vehicle of a national evaluation to investigate summative and formative questions.

We discuss each of these issues in turn.

Social Competence

Zigler (1973) originally suggested the concept of social competence as the criterion of success for Head Start. He elaborated the term as meaning "an individual's everyday effectiveness in dealing with his environment. A child's social competence may be described as his ability to master appropriate formal concepts, to perform well in school, to stay out of trouble with the law and to relate well to adults and other children." Zigler proceeded to identify three components of social competence: a child's health, intellectual ability, and social-emotional development.

Using increased social competence as the criterion of Head Start success raises definitional and measurement problems. The prior and very serious problem is definitional. What is meant by the phrase "social competence?" This term is not a notation for a well-understood concept but has different referents--many of them vague--for different groups in various social contexts. If it is to be used as a criterion variable, however, it must be defined precisely before the question of its valid and reliable measurement can even be asked.

In the philosophy of science, two traditional strategies for defining a concept are *nominal* and *real* definition. We will not deal further with the first kind of definition--nominal definitions are not appropriate for defining a concept that represents the success criterion of a social program. A *real definition*, appropriate to our definitional problems, attempts to single out the essential characteristics of a phenomenon or entity--e.g., A chair is a piece of furniture with a seat for one person.

Hempel (1952) distinguishes three types of real definitions: *analytic*, *empirical*, and *explicational*. All three types had to be used in the Head Start evaluation design. For a variety of reasons, none could be used satisfactorily.

An analytic definition of a term is a *meaning analysis* of that term. It analyzes the meaning of a term into components whose meanings are known. An example of such a definition is Zigler's equation of "social competence" with "an individual's everyday effectiveness in dealing with his environment." Another example derives from the statement of Head Start objectives in the *OCD-Head Start Policy Manual* (p. 7):

A. The improvement of the child's health and physical abilities.

B. The encouragement of self-confidence, spontaneity, curiosity, and self-discipline, which will assist in the development of the child's social and emotional health.

C. The enhancement of the child's mental processes and skills, with particular attention to conceptual and verbal skills.

D. The establishment of patterns and expectations of success for the child, which will create a climate of confidence for his present and future learning efforts and overall development.

E. An increase in the ability of the child and his family to relate to each other and to others in a loving and supporting manner.

F. The enhancement of the sense of dignity and self-worth within the child and his family.

The successful use of an analytic definition presupposes precisely determined meaning and intra- and interpersonal uniformity of usage. It is then possible to appraise whether the two sides of the equation in the analytic definition can be considered synonymous. Neither of these conditions is satisfied by a natural language, to which such terms as "social competence" belong. Although we could not really be successful in determining an analytic definition of social competence, we tried to pursue such a definition as far as we could. We describe those efforts below.

The concept of social competence has meaning only in a social context. Most Head Start children have to deal with two major social contexts, that of the dominant culture and that of their own cultural group. For some groups, the "social competence" effect of Head Start implies ability to deal with the dominant culture. For others, as reflected in the *OCD-Head Start Policy Manual*, it means ability to deal with both cultures. In either context the ability to deal effectively with one's environment entails adequate functioning in a variety of roles. Any valid definition of what social competence means with respect to performance as a family member, for example, or as a member of a specific cultural group or community, or as a peer in a play group or--later--in a working group will clearly derive from the value systems and standards of an individual's family, community, play group, or work group. It follows that there can be no monolithic definition of a "greater degree of social competence," and that changes in the child's behavior must be evaluated in the context of specific roles and value judgments.

The next question is what roles we can expect Head Start to affect--i.e., what roles are available in Head Start, and in which ones might Head Start children be expected to be less competent. Head Start does not offer the child family roles or community roles, nor is there any reason to think these children lack competence in such roles. However, Head Start makes available a *preface* to the role of pupil, and we can expect Head Start-eligible children to be less prepared for that role than non-Head Start children. This role, more than many others, is defined by the dominant culture, and we can expect the Head Start-eligible child to have less opportunity to deal with that culture--which he must do if he is to function successfully in it. His first real experience (as contrasted to vicarious exposure such as through television) is likely to come in school, where he must perform in the role of pupil. The child is expected to assume a new role, learner, under the guidance of a new authority figure, the teacher. The transition period in the Head Start year can be seen as an introduction to these unfamiliar role relationships that makes possible increasing differentiation between roles. There is opportunity to learn about what other people (teachers, parents, peers) expect in the new setting. Head Start also aims to further the skills and behaviors needed to meet the expected role requirement.

Although this argument has not led to anything approaching a precise definition of "social competence," it has allowed us to restrict the term to "effectiveness in the role of pupil." To the extent that the educational system sets up a specific environment with which the child has to cope, "effectiveness in the role of pupil" can be broken down into a number of dependent variables that reflect the child's competence in that environment. The appropriateness of these variables for *all* groups of Head Start children depends on an assumed consensus about what it means to be an effective pupil and about the desirability of playing that role well. Members of different social groups probably vary in how they define and value that role. However, relative to other roles in the society, the role of pupil seems to command an unusual degree of shared definition and value. For example, parents of different socioeconomic classes and ethnic origins tend to want their children to do well in school, in the recognition that the availability of many life options in life depends on being a successful pupil.

If there is no such consensus, a national evaluation using a common set of outcome measures is an inappropriate strategy for assessing the social competence outcomes of Head Start.

The evaluation design also relies on a consensual *empirical* definition of social competence. An empirical definition attempts to define a term as a *function of necessary and sufficient conditions*. Thus, the definition has the status of a systematic empirical relationship. and its satisfactoriness is judged relative to empirical evidence. Although Zigler may have intended to employ an analytic approach, two of his definitions could be considered empirical definitions--i.e., the equation of social competence with "ability to master appropriate formal concepts, to perform well in school, to stay out of trouble with the law and to relate well to adults and other children" and with a "child's health, intellectual ability, and social-emotional development."

In the evaluation design the equation of social competence with the domains of health and nutrition, perceptual-motor/cognitive/language skills, and social and personal development and with specific variables within these domains is an attempt to define social competence empirically. The satisfactoriness of this definition is restricted by the state of child development theory. Although outcomes have been selected for the evaluation that are generally considered important in the development of a child's competence, no one is clear about the necessity or sufficiency of these conditions.

A consequence of our limited empirical knowledge is the inclusion of a fairly large set of dependent variables. This increases the difficulty of carrying out a national large-scale evaluation, further compounded by the lack of appropriate assessment instruments. A second result of using a large number of dependent variables may be that unrealistic expectations are set up for Head Start. If gains are shown only for a small number, we will not know whether this is because Head Start is ineffective or because we have included too many outcomes not likely to be affected. It is unfortunate that the consumers of evaluation information may tend to draw the first inference.

The third type of real definition, *explication*, is a *logical* analysis of a term. This type of definition is concerned with natural language concepts that are vague in meaning, and it attempts to give them more precise meanings so as to render them more suitable for clear and rigorous discourse. Although such a definition usually starts from customary meanings, it usually produces a reinterpretation of terms. It tends to be theoretically elaborated in nature, suggesting that a concept *ought* to have a particular meaning if it is to serve the explanatory function we have specified for it. Some of the recommended measures of social and personal outcomes in the evaluation rely on initial stages of explicational definition: They are based on specifications and extensions of customary meanings of social competence. A careful explication of the concept of social competence is badly needed in the field of child development. However, this is a *major* concept formation task; and it is a difficult one to conduct in the context of designing a national evaluation.

Payoffs from a Longitudinal Study[1]

The initial formulation of the proposed evaluation was a five-year longitudinal study of two cohorts of Head Start children and controls through the third grade. The reason for investing in such an effort would be to establish whether Head Start has long-term effects.

We have two comments here. One involves the expected *substantive* and *policy payoffs* from a longitudinal evaluation. The second addresses the *accountability criteria* for Head Start. When we talk about the longitudinal effects of Head Start, we can distinguish three kinds of long-term effects:

---

[1] We thank John Butler for contributions to this section.

1. Effects that cannot be lost or transformed (e.g., immunity against polio);
2. Effects that can be lost, but are not expected to be transformed as a function of the maturation of the child (e.g., acquisition of certain vocabulary items); and
3. Effects that are expected to be transformed as a function of the maturation of the child (e.g., reading ability).

For the first type of effect, knowing the immediate effect is to know the long-term effect. No longitudinal evaluation is needed for effects of this sort.

Assessing the second type requires a longitudinal design. The effect and consequently the measure are the same across time.

Assessing the third kind of effect also requires a longitudinal design. It requires specifying the *genotypic* dimension that Head Start is expected to affect; specifying the *phenotypic* behaviors that would be expected at different ages, as a function of maturation on the genotypic dimension; and creating measures appropriate to the different phenotypic behaviors. Assessment of this third type of effect poses formidable theoretical and measurement problems. For many dimensions there is little agreement about the presence of a developmental sequence, let alone its nature. If there is knowledge of a developmental sequence or alternative sequences, it is often the case that the behavioral indicators of the dimension are expected to be different at different ages. Thus, noncomparable measures have to be used to assess the presence or absence or nature of the dimension. Differences between Head Start and control groups over time on the dimension can then be attributed to differential reliability and validity of measures, as well as to standard alternative explanations of differences.

In other words, we do not know how to evaluate these kinds of effects of preschool programs, except for academic performance. Longitudinal academic effects are, however, only one part of social competence, and there is already a considerable body of evidence available on them. True longitudinal effects related to social competence would include decreases in adolescent truancy, crime rates, alcoholism and other symptoms of inability to cope; and increases in economic self-sufficiency, family stability, active citizenship, and life satisfaction. Even where indicators for such outcomes are available, there is no theory to let us link them to the Head Start experience.

Even if we have the theoretical and measurement bases for assessing long-term effects, there are other problems. The cost of following up a nationally representative sample for three or four years is exorbitant.[1] If the funds and effort were invested to follow up the children included in the original Head Start and control groups so that attrition could be kept to reasonable levels, many cell sizes would shrink to the point that they could not be used in any analysis. The Head Start catchment areas are not the same as school attendance areas; therefore, children at a specific center (and the corresponding control group) may well go to different schools and have different experiences. In many cases, one would have to pool data across different schools to maintain an adequate cell size, thus confounding differential school effects with long-term Head Start effects, even for a single group of children who were originally exposed to the same treatment. Thus, both the power of any analysis and the inference base are likely to be greatly eroded, with the predictable result that *little credible new information will be obtained at great cost.*

Even if a longitudinal study yields substantively important and interpretable data, the nature of its probable policy payoff should be understood. As Stearns (1973) has pointed out, "the policy decision for which the study is conducted has usually long since been made by the time the study is done." This is patently true of longitudinal studies. For example, if OCD were to proceed with the proposed follow-up into third grade, *results on the longitudinal component of the evaluation would beome available in 1980 at the earliest*. This is an acceptable time frame for research, which can make important contributions of its own to policy formulation, *but not for impact evaluations* expected to yield information for year-to-year resource allocation decisions.

---

[1] Even the successful, if Herculean, effort to maintain the sample for the ETS Longitudinal Study is no guide, since sites were selected expressly for their stability, an unacceptable criterion if representativeness of sample is desired.

The more general question is whether Head Start should be held accountable for long-term effects. If a child exhibits greater curiosity at the end of Head Start, but appears to have "lost" that by the end of third grade, does it make sense to hold Head Start accountable for the loss? If a child "loses" cognitive gains he exhibited after Head Start by the end of third grade, which experience is more plausibly accountable, Head Start or the second grade?

One always hopes to discover that a limited intervention such as Head Start has large effects in the child's life. However, there is no reason to expect this to occur. The more general result in social change programs--e.g., public health programs to persuade Americans to stop smoking or to use seat belts--is that small interventions have small effects. In the case of Head Start, the child has three or four years without Head Start influence before he enters the program. Even during the Head Start year there are several other major influences in the child's life: home, peers, community. If these influences--and the school into which the child enters--reinforce what Head Start attempted to do *after* the Head Start year, then we might expect a long-term effect.

Our position on accountability is that the investment in Head Start is justified if it gives a child multiple ways to develop his skills and use them effectively in meeting the role expectations he encounters at that immediate juncture in his life. These effects of Head Start can and should be assessed at the end of the Head Start year or at the beginning of first grade. They do not require a longitudinal study.

We do recommend evaluation over time, if not in the traditional longitudinal sense. Head Start is a very complicated system; no single impact evaluation can possibly estimate all of its important aspects. The system also alters over time. Thus, it makes sense to think of "evaluation" as *multiple* evaluations and as a *process*. In this conception, studies are done to buy more intensive, extensive, precise, or current information about the state of the system. To this end we have recommended throughout the report and in Chapter 10 a few special studies in addition to the national evaluation.

The feasibility of this conception of evaluation depends heavily on the Office of Child Development. A group of studies is simply that, unless a central group at OCD integrates the results of studies into a system of knowledge about the program and selects new studies on the basis of the implications of that integration.

Evaluation as Multiple-Constituency Process

During periods of this century, scientific work has been regarded as an objective and neutral process. If the scientist and subject were considered at all, they were regarded as objects in the process. Recently an alternative view has emerged in which scientific work is conceived to have social biases and political implications. Scientist and subject are no longer passive elements in the process. Questions about the scientist's responsibilities and the subject's rights have emerged. The older concern for physical or psychological damage to subjects has enlarged to include social damage to groups as a function of culturally biased outcomes and measures. This conception explicitly rejects the view of evaluation as an objective and neutral process.

We therefore assume that the national evaluation of a federally funded social program is a highly political process. As in any such process, there are multiple constituents, each with certain rights and obligations in the process. We recognize the following constituents in the evaluation design and assume that they are equal in claim. The order of listing is not indicative of priority of claim. The constituents are: Head Start-eligible children; the parents of these children; the cultural groups to which the children belong; the personnel responsible for the program at local levels (e.g., Head Start teachers, center directors, program directors); the federal groups responsible for the program (e.g., Office of Child Development, Assistant

Secretary for Plans and Evaluation in HEW); the scientific community; the scientists who conduct the evaluation; and the taxpayer, as represented by the Congress.

It would be preferable to state constituent rights and obligations as a formal system. However, at this point in the history of evaluation research, new constituent rights are emerging and old ones are being redefined. Consequently, it makes more sense to restrict ourselves to stating the major rights we try to recognize in the evaluation design. These are listed below. Some are specific to one constituent; some are common to all constituents.

- Right to know whether money spent on Head Start is being spent well, relative to other uses for that money for poor children (taxpayer; Office of Child Development; Assistant Secretary for Plans and Evaluation; Head Start-eligible children; their parents).
- Right to evaluation within a culturally appropriate framework, including culturally appropriate outcomes and culturally fair measures. (Head Start-eligible children; their parents; cultural groups of which they are members; Head Start personnel).
- Right to freedom from physical, psychological, or social harm in the subject-tester interaction. (Head Start-eligible children; their parents; Head Start personnel).
- Right to confidentiality of data on individual children and individual sites (Head Start-eligible children; their parents; Head Start personnel).
- Right to assert alternative ciriteria for success of Head Start (all constituents, especially Head Start parents and program personnel).

- Right to assert alternative weightings of indicators of the success criterion (all constituents, especially Head Start parents and program personnel).
- Right to meaningful data (all constituents).

A national evaluation of a federal program primarily serves the *information* needs of national, not local, groups--e.g., OCD, the Congress. It is probably impossible to address the information needs of both national and local groups in a single study. However, research conducted primarily for national groups does not have to be done and should not be done at the expense of local groups.

Payoffs from a National Evaluation

Evaluation of a program can be conducted for several different reasons:

1. Assessing the program's worth, where "worth" is defined by some criterion (e.g., increasing social competence of young children)--summative evaluation.
2. Estimating the effects of variations in program inputs on program outcomes in order to improve the program--formative evaluation.
3. Assessing conformity of program operation to program intent--monitoring.

The initial intent of the OCD grant to Rand was objective 1, but as the grant period progressed, OCD also expressed interest in objective 2. Objective 3, with respect to Head Start, is being pursued by OCD through other means and we do not discuss it here. We argue below that we have serious reservations about a national large-scale evaluation of the kind originally envisioned by OCD to meet objective 1. We further argue that the information base available for Head Start does not meet the requirements for addressing objective 2 by means of a large-scale study.

Objective 1. Two sets of questions arise in the use of large-scale national evaluations to assess the worth of the Head Start program. The first set of questions revolves around the probability of obtaining reliable, valid, and interpretable data about program worth, given social competence as the criterion variable. The second set involves questions about why it is important to know about program worth, what OCD can be expected to learn about program worth, and whether or not it is more important to know other things about the program.

There is no point in conducting the evaluation unless it yields reliable, valid, and interpretable data. Several properties of the Head Start case reduce the probabilities that usable data can be collected. First, an estimate of program worth requires estimates for children of different ethnic groups. This introduces what has come to be called the problem of culture-fair outcomes and measures, but it is in fact the standard problem of cross-cultural research. Second, a random assignment design or a value-added design free of or corrected for age selection interactions is necessary if the data are to be creditably interpreted. Third, we expect considerable variations in outcomes within the Head Start program because of variation in the treatment and in eligible children. Thus, the sample has to be fairly large to assure an accurate estimate of effect. Fourth, if we understood the necessary and sufficient conditions for "social competence," we might have a small number of variables or proxy variables that adequately indicated the child's social competence position. However, we do not know these conditions. Consequently, we are thrown back on the variables that seem "relevent" or "important." The minimum number of variables that should be included on these bases is large.

In sum, obtaining an accurate estimate of program worth requires evaluating a large number of variables for a large sample with cross-culturally appropriate outcomes and measures and with a design that has logistic or political difficulties (random assignment) or methodological difficulties (value added). The outcome/measures/design problems are direct threats to the validity, reliability, and interpretability of the data. The number of variables and required size of the sample are management problems, which in turn are direct threats to the validity, reliability, and interpretability of the data. All of these problems *can* be handled. However, to do so will require *substantial cooperation and commitment from all constituents to the evaluation*. It will almost inevitably exhaust the good will of some of them.

Whether it is important to know about program worth is a question only OCD can answer. The usual reason for wanting knowledge of this kind is resource allocation: What does the investment in Head Start of $400 million or more per year yield? This sort of question is appropriate if Head Start is considered: (1) a prototype for a service program to be made more broadly available; or (2) to be in competition with other programs for the same resources. Before we can make either decision, we must know what its effects are. But for that particular purpose, the effects of the program must be construed more broadly than the specific results in child development. Butler (1974) notes that other criteria may well carry higher priority for decisionmaking than gains on measures of child behavior and achievement--for example, whether the program satisfies the constituents for whom it has been designed, or whether it is effective in bringing needed services to low-income families. White (1973) also notes the importance of parent loyalty to the continuation of the Head Start program. Still another criterion, akin to equality of educational opportunity, derives from a consensus view of the multiple options that should be available to young children in this society. A different kind of criterion might be whether the two or three hours spent by children each day in the Head Start center enhance the quality of their daily lives, without regard to how the children might be helped by that experience to cope with later responsibilities in life and school.

OCD is the best judge of what kind of information it needs about the program to defend current expenditures of funds for the program or to argue for increased expenditures. It is important, however, to keep in mind what will probably be learned in a national large-scale evaluation. Let us assume that the data are reliable, valid, and interpretable. OCD will then know how Head Start is affecting eligible volunteer children in general, and by culturally distinct groups, for a number of dimensions believed to affect social competence but not now known to be necessary and sufficient for social competence. Unless we have been successful in our selection of Head Start input variables, OCD will not know what it is about Head Start that does or does not produce an effect for children. The question then arises: What other uses does OCD have for its evaluation funds, and would these be better spent learning about other things? There are limited research funds, limited amounts of OCD monitoring energy, limited scientific energy, and limited tolerance of local program personnel and Head Start parents for measurement of the children. A national evaluation consumes large amounts of all of these resources. *Unless it is imperative that OCD evaluate program worth at this time, these resources are almost certainly better spent in small careful studies that inform us about the necessary and sufficient conditions of social competence, which of those conditions might conceivably be affected by social interventions of any sort, and what kinds of interventions have the desired consequences.*

Objective 2. Estimating how variations in program inputs offset outcomes requires a prior classification of those variations. We expect substantial variation in program inputs. It is not known, however, what variations occur at which Head Start sites. More important, it is not known what variation in programs can be expected to produce variation in effect. A national evaluation conducted for other reasons might shed light on which variations between programs cause variations in effect, but there is not sufficient knowledge to justify a large-scale evaluation primarily for this reason. *The preferred strategy for objective 2 is a series of studies using small samples.* The reason is that since we have insufficient information on the nature of the input variations and on the cause-effect relationships between input and outcome, any evaluation aimed at program improvement is exploratory rather than confirmatory. This produces a high error situation; in such a case it is more efficient to make errors with as small a sample as possible. It is also a multiple-variable, possibly intensive measure, situation. There are tradeoffs between sample size and the number of variables that can be adequately evaluated for that sample. Again, the requirement is for a small sample. The range of variations is likely to require that a number of sites be studied; hence, several small-scale studies will be needed.

Negative Payoffs. Besides posing questions on program worth and program improvement, those in positions of responsibility must ask themselves what the unplanned consequences of a new evaluative effort might be (see Cohen, 1974). Too often, the results of large-scale evaluations asking fairly limited questions are interpreted as sweeping estimates of programs. When programs appear to be "ineffective," unwarranted conclusions may be drawn about the nature of the original social problem or its remedy, and the program abandoned. In the past, these policy positions have often been based more on value judgments than on evaluation results. However, results of evaluations can be used to support such positions. Unfortunately, it does not usually work the other way. By themselves, evaluation results seldom cause a change in value positions--the evidence is never persuasive enough to reverse strongly held opinions.

It is important to recognize that service components of social programs *should*, in fact, embody social value positions. For example, it is well known that nutritional supplements are most effective in preventing later problems when provided at the prenatal and neonatal stages. But this does not mean that hungry four- or five-year olds should not be given hot breakfasts and lunches. The criterion of long-range benefits is irrelevant to our desire to see hungry children fed. Similarly, it could be argued that Head Start incorporates a "quality-of-life" value that society holds for its children without having to prove its merit on the basis of short-term or long-term effects in child development. The need to understand effects of programs is important with respect to improving those programs (objective 2). The effects of such knowledge on resource allocation decisions (objective 1) is often unpredictable, because such decisions are embedded in the policy issues of the day. We have argued that *large-scale evaluation is the wrong vehicle for investigations aimed at program improvement*. Though we have, in the following chapters, provided a design for a national evaluation, *we urge that the need for establishing the worth of Head Start through this means be reexamined in light of the likely positive and negative payoffs.*

In sum, we advise against a national evaluation for either objective 1 or objective 2. Our preference would be that the OCD pursue a system of small studies to investigate these and other questions of concern to them.

Chapter 2

EVALUATION OVERVIEW

OVERALL EVALUATION STRATEGY 32

OUTCOMES 36
  Lack of Adequate Theory 37
  Head Start Effects 38
  Selection Process 40

MEASURES 42
  Problems of Validity 43
  Selection 45
  Recommendations 47

INDEPENDENT VARIABLES AND THEIR MEASUREMENT 48
  Selection 49
  Measures 50

EXPERIMENTAL DESIGN 51

STATISTICAL ANALYSIS ISSUES 53

FOCUSED STUDIES 54
  Test Development 54
  Integrity of the Evaluation 55

Chapter 2

EVALUATION OVERVIEW

Chapter 1 attempted to develop some guiding principles for the proposed evaluation. The principles derive from an analysis of the history of past evaluations and the assumptions underlying the new evaluation as originally proposed. A major theme that emerges is the importance of understanding the conditions that make a national large-scale evaluation the design strategy of choice. Though we are skeptical that all the necessary conditions are likely to obtain, we have provided, in Chapters 3 through 9, the basic design elements for such an evaluation. Throughout the text and in Chapter 10, alternatives are suggested in the form of small-scale, focused studies. A major decision for OCD is whether most of its evaluation resources should be invested in one large-scale impact study or in a series of small-scale studies addressed to specific policy questions.

In the remainder of this chapter we discuss the detailed reasons for choices made with respect to the main elements of the design for the evaluation:

- Overall evaluation strategy.
- Outcomes (dependent variables).
- Measures to assess outcomes.[1]
- Independent variables likely to influence outcomes.
- Measures to assess independent variables.
- Experimental design.
- Statistical analysis issues.
- Pilot evaluation.

OVERALL EVALUATION STRATEGY

The research questions addressed by the evaluation are:

---

[1] "Measures" is used as a generic term encompassing the full range of assessment techniques.

1. What are the social competence effects of Head Start for members of the eligible population who receive the treatment, relative to members of that population who do not?
2. What are the social competence effects of Head Start for eligible children from different cultural groups who receive the treatment, relative to eligible children from those same groups who do not?
3. What are the social competence effects of Head Start for eligible children within each cultural group who receive the treatment and who differ in entry characteristics, as indicated by pretests and other background characteristics?
4. Are there any indications that variations in treatment produce variations in social competence outcomes for children who receive the treatment?

The recommended strategy includes evaluation at three levels:

- A large-scale study to investigate the effects of Head Start on a representative sample of the Head Start population;
- Subsamples drawn from the large-scale study population for the administration of special measures not suitable to the entire sample or infeasible for large-scale administration;
- An independent series of focused smaller studies to address specific questions that cannot be answered in the context of a large-scale study. If this series of studies is to provide information relevant to future decisionmaking, an analytical methodology will have to be developed for tying together results from all three of the evaluation levels.

The first two levels (large-scale study and subsamples) make up the basic evaluation. Since the major question to be answered by the large-sample evaluation is one of overall Head Start effects, the sample should be broadly representative of the Head Start population, and the test battery should address common goals. Major sampling considerations include ethnic classifications (Black, White, Puerto Rican, Chicano, Native American), size of community (metropolitan/nonmetropolitan), and regions of the country corresponding to cultural distinctions within ethnic groups.

Subsamples will be drawn for two reasons: (1) where outcomes are sufficiently important that they need to be measured, but appropriate tests are too costly or complex for administration to the whole sample; and (2) where different measures must be used to capture the same outcome for different population groups--for example, in language development for Spanish-speaking children. The first reason requires subsamples drawn randomly; the second requires knowledge of variations in the expression of behaviors and skills in different population groups and the availability of measures that are valid for those variations.

The choice of focused, small-scale studies should be made by identification of policy-relevant questions in three areas: (1) methodology, (2) substance, and (3) experimentation. The purpose of *methodological studies* is to provide information or techniques instrumental to carrying out some aspects of an overall evaluation plan. For example, separate studies are necessary to identify the outcomes to be included in succeeding stages of a continuing effort. One such study might combine survey, structured interview, and environmental observation techniques to arrive at operational definitions of social competence goals within families and within the various cultural groups served by Head Start. These definitions could then be used to construct instruments to assess the child's state on family and community competence outcomes. A methodological study of a different kind is the development of quasi-Bayesian models that would allow the cumulation of results from past evaluations and from the proposed three-level effort.

We define *substantive studies* as investigations of effects of naturally occurring phenomena within the settings being investigated. Such studies should derive from policy questions that cannot be addressed in the context of large-scale samples. One example is the investigation of Head Start effects on personal and social development as evidenced within the home, a project that would use the results of the methodological study noted above. Another example is the series suggested in Chapter 1 on differential long-range effects of Head Start depending on whether school pedagogy and philosophy are congruent with the preceding Head Start experience or discontinuous with it. This kind of study requires detailed monitoring of both the Head Start and the school experience, feasible in only a few instances to be selected to represent the widest possible contrast. A third example of a set of substantive studies is the investigation of differential effects of the new Improvement and Innovation program alternatives. At this time, the instances of each variation (home-based, fewer-than-five-days per week, etc.) are few compared with the number of Head Start centers, and their operational characteristics are much less well known than those of the "traditional" five-day center-based program. Hence, their inclusion in the sample for the large-scale evaluation is inappropriate. For program planning, however, it is important to compare the differences, if any, in effectiveness between these variations and the more traditional mode of operation.

Experimental studies involve investigations of effects of manipulated (i.e., created) interventions. They are aimed at program improvement. Studies would center on attempts to introduce specific new program elements in order to investigate what important outcomes can be taught and, if attempts are made to teach them, in what ways resulting effects can be measured. For example, the teaching of metalinguistic and metacognitive skills and their assessment are at present in the domain of basic research. But since these appear to be important skills for functioning in a variety of roles, some resources should be invested to see whether they can be taught in the Head Start setting and to develop adequate assessment techniques. The OCD should consider the effect of providing evaluative feedback to Head Start teachers on a continuing basis so as to improve their ability to meet the needs of individual children (a suggestion made by panelist Scarr-Salapatek and endorsed by the Panel on Cognitive Development).

Chapter 10 will discuss some specific focused studies in greater detail.

OUTCOMES

The objective of OCD for the proposed evaluation is to assess whether Head Start increases the social competence of disadvantaged children. Outcomes concerned with direct effects on parents, communities, or service agencies are not to be included. In Chapter 1, we have defined the concept of social competence in role-theoretic terms. This approach, using the traditional areas of child development as the knowledge base, asks the question: What physical attributes and skills, sensory-motor and perceptual abilities, conceptual base and communication skills, and personal and social attributes does the child need to function in the various roles appropriate to specific environments? A second question in selecting outcomes is: For which of these attributes, behaviors, and skills is Head Start likely to make a difference? It would be an inefficient use of resources to assess dimensions that Head Start has neither the purpose nor the power to affect. Third, if we expect to find any difference made by participation in Head Start, we must also be concerned with the presence of variation between children on a particular behavior or skill, and with the number of children for whom a change is sought. Thus, the major criteria for selecting outcomes are:

1. *Importance*. How important is the specific characteristic behavior, or skill to the child's social competence? Does its absence or "inadequate" presence interfere with the child's functioning at a level to be expected for his age?
2. *Variation*. Is this a dimension of behavior or performance on which children show variation--i.e., do children have room for movement on that particular dimension?
3. *Head Start effect*. Does Head Start try to help the child acquire or strengthen the characteristics, skill, or behavior?
4. *Occurrence*. Is the incidence of need or inadequacy that may create a problem for the child sufficiently frequent that change, when brought about by Head Start, would be apparent in an evaluation with limited sample sizes?

Lack of Adequate Theory

A number of problems impede the application of these criteria in a rigorous fashion. The first is the absence of a comprehensive and generally accepted theory of child development from which could be derived characteristics, competencies, or behaviors prerequisite to some succeeding desired set, such as attitudes and skills effective in the context of school or family. This makes the criterion of *importance* difficult to apply.

The problem is eased somewhat by conceptualizing each major area of child development somewhat differently in accordance with its underlying theoretic framework. The differences are elucidated by several of the models recently posed by Mercer (1974). The first, the medical model, is derived from physiology and tends to concentrate on defining the nature of the abnormal and its etiology. The model assumes value consensus and focuses on the individual apart from his sociocultural setting. Although health and nutrition status is not independent of sociocultural setting, its assessment is concerned with basic biological processes that operate similarly in all human beings. Findings can be generalized with a high level of validity transcending social system boundaries. Hence, this model is more appropriate than others to the area of health and nutrition where we wish to assess the effects of Head Start on the absence of disease and deficiencies and on the presence of good health.

A second model, the pluralistic model, derives from the traditional statistical model based on the normal distribution curve. Unlike that model, however, it does not use unitary norms that lead to erroneous generalizations over dissimilar populations, and it focuses on what a person has learned rather than on aptitude or intelligence. This model is appropriate to performance in the perceptual-motor, cognitive, and language skills area. It implies performance-based tests and within-group comparisons.

The third model proposed by Mercer, the social-system model, is concerned both with general behavioral expectations for anyone who is part of the system and with specific role expectations that the individual has to meet. It thus shifts the focus from the individual to the normative structure of the social system by which his behavior is defined. "How behavior is defined; the types of social system in which the child lives; and the roles which he plays in these social systems are central concerns" (Mercer, 1974, p. 14). This model provides us with an overarching framework for any evaluation of social competence and is particularly relevant to social and personal development.

Head Start Effects

A second problem in selecting outcomes derives from the application of the third criterion, the likelihood of a *Head Start effect* with respect to a specific outcome. We have made the point that the interactions between child characteristics and progam characteristics are often crucial in considering overall outcomes for individual children. But past analytic strategy has generally emphasized a definition of effect that focuses on average gains for average children in average programs. These averages are numbers without much meaning, since one number cannot mirror the properties of such a complex system as a child or a program. Since most social interventions are highly variable in their site-specific implementation (especially if, as in Head Start, local control and decisionmaking are mandated), the averaging over programs and children can serve only to wash out specific effects that might indeed be present for individual children in specific programs. Hence, past evaluations do not provide us with much information on what important outcomes Head Start is likely to affect outside of the documented short-range cognitive gains.

The basic premise of the program is that children from low-income families can draw particular benefits from a child development program. In what way is the premise likely to be correct? In the health and nutrition areas, all available evidence points to the fact that children from low-incomes families do not on the average receive the nutritional and health care benefits available to middle- and high-income children. Head Start attempts to remedy this by providing hot meals, snacks, and a program of medical care involving preventive measures, diagnostics, and follow-up treatment. It is therefore appropriate to select outcomes than can serve as indicators of a child's medical and nutritional status to assess whether participation in Head Start leads to improved health.

Head Start also attempts to improve the social competence of the child. But there are no *a priori* reasons to assume that a low-income family is not as competent at socializing the child to his family responsibilities as the family with a higher income. Nor is there any reason to assume that the cultural group to which he belongs is any less competent than any other at teaching the child the specific competencies needed as a member of that cultural group. It is therefore unclear in what ways a developmental program could be differentially beneficial for the low-income child as compared with the middle-class child with respect to his role as a family member, a member of a specific cultural group, or a member of his peer group.

A child from a low-income family may well have had less incentive or opportunity to learn the competencies necessary to cope with school, since that institution embodies, by and large, the values of the middle class. Therefore, if coping with school is accepted as a consensual goal for all children, then a relevant function of Head Start is to help smooth the child's transition into the school environment. Because the setting is explicitly designed to accomplish it, and because the child's needs may be greatest in that area, one would expect significant effects of Head Start on the child's competence in functioning in the role of pupil.

The arguments made above provide a decision framework for selecting outcomes appropriate for a representative population sample of Head Start children and those outcomes that must be considered specific to a given setting. In the first category are outcomes that represent consensual goals in health and functioning in the school environment; in the second category are locally defined changes in competence with respect to functioning as a member of the family or the community. The first leads to a common set of outcomes and measures appropriate to a national evaluation and the second to variations in outcomes and ways of assessing them consonant with values of the community. Further, there is considerable knowledge on what constitutes good health and nutrition (or at least the absence of problems) and adequate educational performance, but little information on effective functioning in other areas. Even if we could identify some relevant outcomes, their measurement often presents problems of intrusiveness that can be overcome only by highly costly anthropological fieldwork. Thus, a considerable amount of exploratory work must precede any attempt to assess the possible influence of Head Start on the child's social competence in relation to family and community.

Divergent goals on the part of participants and staff are also related to the possible effects of Head Start. For example, it may be important to know teacher's ratings of the selected outcomes as well as their actual emphasis (as gauged by their day-to-day classroom behavior), since this may well change the character of the experience the child undergoes. Analogously, the goals of parents, the community, and the federal and local sponsors will all operate to shape the nature of the program, with consequent changes in important program variables. Therefore, *we recommend that Head Start personnel and clients be given an opportunity to assign their own weightings to the major areas measured by the test battery, and be given their choice of two outcomes of their own to add to the basic battery for assessing some local goals* (for example, knowledge of songs and dances indigenous to the cultural groups). This is in accord with the principle discussed in Chapter 1 that evaluative effort must yield information of importance to each legitimate interest group. The assessment of divergent goals requires not only inclusion of site-specific priors and goals, but also appropriate analytic strategies--for example, considering each site as a separate experiment (see Chapters 7 and 8).[1]

Selection Process

Selection of outcomes has been structured both by the state of the knowledge base in each area of child development and by the somewhat different conceptualization of each area described above. The knowledge base is fairly well established in the health and nutrition area, somewhat less so in the area of perceptual, cognitive, and language development, and is least definitive in the area of social and personal development.[1]

---

[1] Guttentag (1973) has developed a model that ties the preferences of central decisionmakers to the data-gathering process. We are suggesting analogous procedures here to take account of the preferences of clients and local personnel in both the data-gathering and the analytical phases.

We chose a three stage method for identifying candidate outcomes:

1. Advice of consultants in health and nutrition, motor and perceptual development, cognitive development, language development, and socioemotional development. This activity was intended to cast as wide a net as possible so that all outcomes that could reasonably be hypothesized would be considered. The search for outcomes in the areas presumed to exhibit common patterns across cultures (physiological, motor and perceptual, cognitive, and communicative processes) involved expert panels supported by commissioned issue papers, written responses to panel deliberations, and individual consultations. For the culture- and role-specific outcomes, convening a panel proved inappropriate because of the several extant schools of social and personality development, so we had to rely to a greater extent on consultation with active investigators and on critical assessment of current research. (For a list of all panel members and individual consultants, see Appendix A.)
2. Examination of outcomes and measures used in previous Head Start evaluations and related projects and, where available, the results of consequent analyses. This provided an experiential base for judgments on what child development outcomes

---

[1] The differing states of the knowledge base are reflected in the somewhat different styles of the chapters dealing with the three major areas. An increasing amount of justification for selected outcomes and detail on measures are given from Chapters 3 through 5. Also, since the area of social and personal development is at the heart of social competence, a more comprehensive discussion is given in Chapter 5, including more suggestions for test development than for the other two areas.

had been considered important in the past and how Head Start had affected these outcomes.

3. Consultation of the relevant child development literature to pursue the suggestions of panelists and consultants, uncover possible omissions, and guide the application of the four criteria used to make selections. This step was particularly important in the areas of personal and social development, where the views of experts as to the significant outcomes to be assessed diverged to a much greater extent than in the other areas, and where experience with findings from earlier studies proved very disappointing.

This three-stage process resulted in a large list of candidate variables, which was then reduced to the list of outcomes included in the basic battery by application of the criteria of importance, distribution characteristics, and Head Start effects. Some candidate outcomes, though considered important, are not included in the basic battery because they must be examined in a more exploratory fashion by means of focused studies. The rationales for specific choices within each area of child development are given in Chapters 3 through 5. The selected outcomes are summarized in Table 2-1 at the end of this chapter.

MEASURES

We have already noted some of the difficulties with existing and previously used measures. These difficulties expanded the task of selecting appropriate tests and measures beyond the simple matching of instruments to the outcomes of interest. It was also necessary to consider what adaptations would need to be made, what experimental techniques should be piloted, and what tests should be developed *de novo*. Four major criteria were applied to the selection of tests and measures:

1. *Validity of test content.* This criterion implies judgments on the degree to which the test tasks to be performed by the child or the attributes or behavior components selected for

rating or observation are congruent with the outcomes they are intended to assess.

2. *Validity for the sample population.* Particular concerns here are age appropriateness and language difficulties that interfere with nonlanguage outcomes. As part of this criterion, cultural appropriateness of any test must be considered.
3. *Technical reliability.* Where possible, instruments should be chosen that minimize variation associated with the measure but are unrelated to its substantive intent. This implies such test properties as high inter-tester reliability and robustness with respect to uncontrollable but irrelevant field conditions.
4. *Unobtrusiveness and reasonable cost.* Tests and instruments that can be administered by community paraprofessionals are preferred over those requiring highly trained outsiders; simple measures are preferred over complex ones, and group administered measures over individualy administered ones. These kinds of instruments reduce the obtrusiveness and cost of measurement.

Problems of Validity

Again, the application of several of these criteria raises problems. In the health and nutrition area, there are few validity problems for the specific measures that assess such outcomes as gain in the child's height and weight, or hemoglobin level. (Even in this domain, there may be problems of uniform administration; for example, whether the child's height is recorded upright or lying down.) The question of validity applies more to the *set* of tests--that is, whether the total subbattery is an adequate indicator of good health and nutritional status, and not simply a source of information on the presence of absence of specific health or nutrition problems. The question arises because although we have excellent measures for very specific outcomes, no broadly valid measures (with the possible exception of a thorough physical exam) are available for indicating good health and good nutrition *in toto*.

For most attributes and skills in other domains, particularly those that are assessed through instruments requiring the child to perform a task, the relationship between the test and the outcome to be assessed is not as clear, nor is there usually a one-to-one correspondence. For example, measures designed to assess perceptual-motor skills may often require a modicum of language proficiency and at least some rudimentary cognitive constructs. In particular, the IQ and vocabulary-based test batteries used in the past tend to be gross proxies for a wide range of skills and behaviors.

Past practice has been to bypass these difficulties by reducing most assessments to *comparisons to age-level norms*, the assumption being that such norms have functional consequences. Comparisons to national norms appear more acceptable in health and nutrition than in other areas of child development, but even in this domain we do not always know the degree of variation tolerable before functional effects occur. In other areas, the assessment criterion of "bringing up to norm" often embodies the notion of lags or deficits in relation to those populations on whom the norms have been based. The questions raised by this procedure are: To what degree are the child development norms based on middle-class populations appropriate yardsticks for the social competence of children from low-income families or from minority groups? Do such norms reflect cross-cultural sequences in the development of social competence of all children or culture-specific behaviors? Standard norms are clearly inappropriate to the assessment of social competence as a family member or as a member of a specific ethnic or cultural group, since coping styles and approved social behaviors vary considerably from family to family and from group to group. In the case of the child's ability to deal with societally defined environments (e.g., school), the generally used norms for performance tend to be congruent with school and teacher expectations. We may indeed be able to identify a *common set* of attributes, behaviors, and skills important to all children. But it is not at all clear that the *identical performance* has the same meaning for different groups of children.

As an alternative to comparisons with norms, one might look for *threshold- or criterion-referenced performance*. Here, as in health and nutrition, there is empirical information. For example, we have available some predictor variables that correlate well with a child's later ability to take advantage of learning experiences offered in school and with his ability to relate to peers and authority figures inside and outside school. But bringing the child's performance up to the indicated level through deliberate instruction may well interfere with the predictive validity of the criterion performance. In that case, achievement attained through successful training on a predictor variable can no longer be used to predict later performance.

A third alternative is simply to note greater *movement in a positive direction* on the part of the Head Start children as compared with the control group. But this kind of assessment technique poses problems of interpretation. What is a "positive" direction? How much movement is necessary? This approach also creates design difficulties, which arise out of having to establish adequate, fully comparable control groups.

Selection

The selection of measures of health and nutrition outcomes is straightforward. Most of the outcomes are associated with standardized medical tests.

Previous evaluations of Head Start have revealed problems with instruments in the perceptual-motor, cognitive, and language areas. Hence, development of a new set of tests was begun by ETS to investigate a number of outcomes for preschool education generally deemed important. The result is the CIRCUS battery of short, easy-to-administer tests for the perceptual-motor, cognitive, and language domains. The battery consists of 14 tests to be administered on a selective basis to children and three instruments that rate testing behavior and educational context. The tests are currently under revision, and validity and reliability data are being analyzed. The whole battery, complete with administration and scoring instructions, will be available by the time the proposed evaluation is to be started. Technical information on test properties will also be made available.

We have selected several tests from the CIRCUS battery that appear to assess outcomes of interest in the perceptual-motor, cognitive, and language domains. Given the availability of a preparatory year for piloting the entire test battery, we strongly recommend their consideration. The rationale behind the development of these tests is more consonant with the purpose of the proposed evaluation--than that of most previously used instruments--that is, to assess the skills developed by young children through a preschool experience.[1] The outcomes included in CIRCUS were selected and test items developed based on the experience of teachers and the research knowledge of child development specialists. Hence, the test battery has a very different genesis from that of most others available (e.g., IQ batteries, clinical diagnostic tests, or normed achievement tests). The new battery is easy to administer and score and holds children's attention, thereby minimizing field problems.

The assessment of personal and social development requires a different approach than does measuring specific attributes, as in health or achievement on assigned tasks. Much of the past assessment in the socioemotional domain has hypothesized the development desirability of certain traits. The definition of the traits and their dimensions have varied with the theory from which they were derived. But results from this approach generally have been disappointing. The relationships between traits and effective functioning are unclear, and failure after failure has been experienced with instruments designed to measure specific traits (see Walker, 1973). Instead, we argue that assessment should focus on situation-specific behavior.

The best available testing technology for this purpose is a combination of rating instruments and structured observations. (Structured observations sometimes involve assignment of a specific task to a child, but the point is to observe his behavior while he is engaged in the task, not to measure his achievement.) Validity and reliability criteria are difficult to meet for these assessment techniques, although there is a considerable empirical base to be drawn upon. In this domain, however, there is very little choice. The one-to-one child-tester situation used in other techniques changes the social context in which the child operates; therefore, it is likely to alter the behavior being assessed and defeats the very purpose of assessing social competence.

---

[1] This is also the purpose of the Preschool Inventory (PSI), but this battery is not recommended because of its cultural bias; i.e., it does not meet the criterion of validity for the sampled population.

Recommendations

The three chapters on outcomes and measures present many recommendations (indicated in the text by italics) and test selections. Most of these concern the measures to be included in the basic evaluation, either for the total sample of children or for subsamples of the total sample in those cases where measures are either inappropriate or infeasible for the total sample. However, there is also considerable discussion of important outcomes that can be assessed only in separate, focused studies, or for which test development is considered sufficiently exploratory that inclusion in the basic evaluation is unlikely. For the reader's convenience, we summarize the basic battery in Table 2-1. Only those measures are listed that are ready or can be expected to be ready by the time they are to be used.

Although measures in the health and nutrition area and in the perceptual-motor, cognitive, and language area are to be administered during the Head Start year, the measures in the social and personal development area are for the most part to be given at the beginning of the first public school year. That is the first occasion on which both the Head Start children and the control group children are being asked to play the role of pupil. This fact also rules out pretests except for those tests being administered during the Head Start year. Some measures in the social and personal subbattery are suggested for repetition, either during the same year or a year later, so that more than one set of data points can be obtained. This subbattery also contains more instruments than the other two that need piloting or further development work, much of which is suggested in Chapter 5 in detail. The reason for needing more preparatory work is the poorer state of the art in this domain; but completion of the suggested tests in time for the evaluation seems feasible since, for most of them, two years of development and pilot time are available.

INDEPENDENT VARIABLES AND THEIR MEASUREMENT

Chapter 6 identifies the major independent variables likely to explain variations in outcomes between treatment and control children and within each group. There are three classes of these variables:

- Treatment variables, such as the kinds of activities children engage in, the time spent in each activity, and the conditions and structure imposed on the activities.
- Variables that specify conditions obtaining for the control group.
- Background variables for treatment and control children, their families, Head Start teachers, the Head Start center, and the site.

The most difficult area in the selection of independent variables is the identification of those variations in program elements (treatment variables) likely to cause differential effects. This is critically true in the area of perceptual, cognitive, communication, and socioemotional processes for which the guidelines given in the OCD-Head Start *Policy Manual* are not sufficiently specific to engender homogeneity among Head Start centers sampled for the evaluation. On the contrary, since local program development is mandated, one should expect variations that could affect the outcomes for children. Since our understanding of the interaction between Head Start and competence development is a crude and incomplete one, we must rely on empirical evidence to identify those program attributes expected to make a difference in the outcomes being assessed. A large number of classroom process variables have been examined in the past, but the yield with respect to the analysis of their differential effects has not been very rich. Even in settings more explicitly structured than the average Head Start center (e.g., HSPV), the identification of critical program components and their stable maintenance have been thorny problems. The criteria applied to selecting classroom process variables are:

1. *Effects*. How likely is it that the candidate variable will affect some of the outcomes selected for assessment?
2. *Likelihood of variation*. Would one expect variations among programs with respect to the classroom process variable under consideration?
3. *Feasibility*. If variations are likely to be present, can they be measured?

Selection

The first stage in the selection of treatment variables was to define Head Start program characteristics and their distribution among Head Start centers, based on an understanding of how Head Start goals are implemented through operations at the national, regional, and local levels. This effort did not yield an adequate basis for formulating treatment variables because data on key descriptive dimensions were not available.

The second stage consisted of a careful review of the theoretical and empirical literature related to aspects of preschool programs that seem to influence child outcomes in those areas in which Head Start specifically claims to intervene. Of special interest were *curriculum* and classroom *process:* Major curriculum variables are educational objectives and instructional strategies for achieving them; process variables include behaviors of teachers and children and educational setting. The voluminous literature on child outcomes as a function of classroom curriculum and process was difficult to interpret because sources used different descriptive systems for presenting classroom factors and different methods for measuring their influence (see Rosenshine, 1971). Thus, it was not clear whether the varied and often inconsistent results obtained were reflecting program differences, procedural differences, or both.

Consequently, a third stage in the selection of treatment variables was to solicit the help of a panel of experts in the field of classroom influences and their measurement (see Appendix A). This group was asked to provide feedback on our preliminary choice of treatment variables and on their relevance to a national Head Start evaluation research design. The final selection was based on the panel's recommendations. The panel emphasized that variations in output is significantly related to variations in educational input, so that dependent variables can be expected to show effects only to the extent that Head Start programs have focused on their enhancement.

Selection of control group variables and background variables was fairly straightforward. We constructed a taxonomy of treatment conditions likely to obtain for control group children on which information could be obtained through a questionnaire. This should give some help in the interpretation of comparisons between Head Start children and controls. To what degree alternative care arrangements also offer the kinds of health and educational services provided by Head Start cannot be determined solely by questionnaires, however. Therefore, it may be advisable to conduct some small-scale studies on the nature of various settings that control children may be in, such as day care centers or informal group care.

Background variables selected on children, their families, teachers, centers, and communities are those that have been found to be associated with variations in outcomes in the Head Start Planned Variation Study or the ETS Longitudinal Study.

Measures

Recently, a number of observational techniques have been developed to characterize preschool or Head Start environments (see Brandt, 1973). Several of these instruments were developed for the Planned Variation experiments in Head Start and Follow-Through and were therefore intended to monitor the implementation of specific models developed by individual researchers. Similar instruments have been developed for closely controlled laboratory studies of preschools. Hence, the applicability of existing observation and rating techniques to characterize the general Head Start setting is not entirely clear. *Pilot testing of the instruments suggested for treatment variables in Chapter 6 is a necessity to establish whether they capture the selected dimensions of the classroom process, and whether the Head Start centers to be included in the sample will actually exhibit variations along these dimensions.* Analysis of the pilot data should give information on whether variations along these dimensions actually cause sufficient difference in outcomes to make their inclusion in the evaluation worthwhile.

Control variables and background variables can be obtained cheaply and effectively through questionnaires and archival information. Table 2-2 summarizes the independent variables selected, the instruments suggested for their measurement, and the time of administration.

EXPERIMENTAL DESIGN

We noted in Chapter 1 that past evaluative efforts have been criticized for their ambiguity and lack of explanatory power. *Any new evaluation effort will be a waste of resources unless it establishes the causal linkages between Head Start and any observed effects with greater certainty than has been the case in the past.* For that reason, the experimental design to be proposed for the new evaluation presumes random assignment of children within a Head Start catchment area (site) to the treatment and control conditions.

The Head Start program imposes important constraints on a design. First, we expect considerable variations within each treatment level--i.e., within Head Start and within the control condition (all alternatives to Head Start). Even if the variations can be distinguished, they cannot be randomly assigned to sites. Thus, these variations are almost certainly confounded with sites. This precludes our obtaining an unbiased estimate of treatment effects across sites unless: (1) children are randomly assigned to control and treatment conditions within site, (2) effects are estimated separately for individual sites, and (3) the effects are aggregated across sites.

A second constraint is that children cannot be randomly assigned across sites. We can then expect confounding of site effects and child background characteristics.

Since the design cannot disentangle variations within treatment and control conditions from site effects and child characteristics from site effects, it is not structured to yield *a priori* estimates for questions 3 and 4. These questions can still be evaluated but in *ex post facto* ways.

The design chapter, Chapter 7, treats these major design elements: the overall rationale for sampling decisions, sampling of treatment, sampling of children and their assignment to treatment and control conditions, stratification dimensions for treatments and children, and sample sizes for site sample and for children per site.

We recommend a two-stage cluster sample of Head Start centers and classrooms within centers. Systematic sampling is advised for selecting centers; random sampling, for selecting classrooms within centers. We recommend selecting children from those who volunteer for the program. Of eight alternative designs for creating treatment and control groups, *we strongly recommend random assignment of volunteers to the treatment and control groups within a site and random assignment of members of the treatment group to the Head Start classrooms.* We explore some properties of the value-added design, recommending it for use in catchment areas where all eligible children in the area are served by the Head Start center. This design is also considered of possible use in centers that either have insufficient demand for the treatment to create a control group by random assignment, or refuse to allow random assignment.

We are unable to specify treatment levels beyond treatment (Head Start) and control (all alternatives to Head Start) conditions. It is hoped that the variables specified in Chapter 6 will distinguish variations within the two conditions in the pilot test of the evaluation.

For statistical and substantive reasons we advise stratifying the sample of children by demographic properties of catchment areas. Ten strata are selected along three dimensions: ethnicity of children served by the center, density or sparseness of population, and region. The ten are: Black central city, Black nonmetropolitan, Puerto Rican, Chicano central city in the Southwest, Chicano nonmetropolitan in the Southwest, Native American with frequent contact with SMSA, Native American with infrequent contact with SMSA, White Appalachia and Ozark, White central city, and all other White nonmetropolitan.

The chapter concludes with a discussion of considerations that determine the optimal sample sizes for the site sample and for number of children per site.

STATISTICAL ANALYSIS ISSUES

All the compromises and decisions with regard to outcomes, measures, design, and data collection become visible in the analysis stage. The analysis history of Head Start is characterized by often ingenious attempts to reverse generally irreversible problems created in earlier stages. The evaluation cannot be rescued at the analysis stage. Specifically, attention should be directed to the following points:

- If there is no clear sense of what a specific score or observation on a selected measure means *prior to the data collection*, analyzing the data rarely imparts meaning.
- If measures are unreliable, administering them reliably does not increase their reliability.
- If measures are reliable, administering them unreliably reduces their reliability.
- If data are unreliable, the variances go up; the higher the variances, the less the ability to detect treatment effects.
- If measures are not comparable, data from them cannot be compared.
- If children are not randomly assigned to treatment and control conditions, differences between conditions cannot be credibly attributed to the treatment.

These points are obvious and may seem gratuitous; however, they can be overlooked in large evaluations. In these cases, a good analysis can make the best of the situation, but cannot redeem it.

It is not the point of the statistical analysis chapter to specify statistical descriptors and tests. It is assumed that the statistical analysis group will know those tests at least as well as we do and the properties of the data better. The point is to explore statistically defensible and meaningful ways to analyze evaluation data, specifically Head Start evaluation data, and to communicate them to different constituents of the evaluation.

The issues addressed are: confirmatory versus exploratory data analysis, hypothesis-testing versus confidence intervals, form of the hypothesis, levels of significance and confidence, interpretation of the results, models for the analysis of random assignment and value-added designs, and aggregation of the measures. The analyst usually has several statistically appropriate solutions for each of these issues. Within the statistical constraints the choices should be based on policy and theoretical considerations. For example, a confidence interval for a mean difference between treatment and control groups implicitly tests a whole range of hypotheses about that difference--i.e., there is no *statistical* reason to perform a test of a hypothesis in addition to calculating a confidence interval. However, for *policy* reasons we recommend that a test of significance and confidence interval be applied. There are standard forms of the null and alternative hypotheses. However, we suggest that Neyman's criterion be used to choose a form and that the application of that criterion be consistent with the Head Start policy environment. As a final example, there are a variety of statistically defensible ways to aggregate the results for separate sites to determine cross-site effects of Head Start--e.g., mean effects, frequency distribution effects, proportion of centers that satisfy increasingly stringent criteria for effect. Of these ways, which is the most meaningful to different consumers of the results? These are the kinds of questions that should be asked if the policy reasons for the research are to be reflected in the analysis.

FOCUSED STUDIES

There are two major purposes of the preparatory period discussed in Chapter 9: to carry out the needed test development and to test out the full-scale evaluation.

Test Development

We urge that test development be carried out by qualified researchers and specialists who are knowledgeable not only about instrument design and construction, but about attributes and the meanings of behaviors within the cultural groups being served by Head Start. Each adapted or newly developed instrument must be adequately field tested and should be used in the evaluation only if it passes the validity and reliability criteria stated earlier in this chapter.

Integrity of the Evaluation

Although there are several criteria that determine the caliber sought for the evaluation, Chapter 9 argues that three are nonnegotiable:

1. *Data quality.* Measures must yield reliable and valid data in the field for all groups of children.
2. *Data interpretability.* Random assignments must be fully explored and instituted where feasible.
3. *Protection of local constituents.* Children, their parents, and their communities or cultural groups must be safeguarded from biological, psychological, and social harm.

The pilot test of the evaluation is designed to determine whether and how these criteria can be met for the full-scale evaluation. As sample sites are selected for piloting the test batteries, means of communicating and coordinating with parents and center staff must be instituted. Objectives and procedures of the evaluation must be explained and local input solicited. In anticipation of the full-scale evaluation, the preparatory year must be used to establish community contacts, appoint local field coordinators and local testers, hold informational meetings on the nature and purposes of the tests in the battery and on the utilization of results, obtain input on importance to the locality of the different outcome areas included in the base battery, and agree on the measures to be used for assessing any outcomes to be added by the community. Random assignment strategies that communities perceive as fair need to be worked out.

Actual field operations during the preparatory period include (a) the pilot test of procedures to be conducted and measures to be administered in the Head Start year; (b) preparation for Head Start year pretests for the full-scale evaluation; and (c) the pilot test of procedures to be conducted and measures to be administered in the post Head Start year. (Phase (c) will actually run into the first year of the full-scale evaluation.) During these field operations adequate quality control mechanisms must be established.

Accurate records of the cost of test administration should be kept and field problems meticulously documented. Before the national evaluation is launched, OCD should consider the information on costs and field problems and preliminary results from the pilot test in light of likely payoffs discussed in Chapter 1. If the decision is made to proceed, the experience of the pilot evaluation should be used to adjust the design for the full-scale evaluation so as to maximize the chances of achieving its purposes.

Table 2-1

BASIC BATTERY--OUTCOMES AND THEIR MEASURES

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
I. Health and Nutrition				
A. Reduction in disease vulnerability				
1. Reduced incidence of disease	Immunization records; TB skin test	W SpS	End HS year End HS year	
B. Reduction in nutritional deficiencies				
1. Iron	Hematocrit	W	(Beg.)/end HS year	
2. Protein	Serum albumin	SpS	(Beg.)/end HS year	
3. Improved dental health	Dental examination	W	(Beg.)/end HS year	
C. Remediation of sensory impairment				
1. Visual	Snellen test for visual acuity	W	(Beg.)/end HS year	
2. Hearing	Pure-tone audiometric screening test	W	(Beg.)/end HS year	
D. Presence of good health and nutrition				
1. In-depth health assessment (optional)	Physical examination	RS	(Beg.)/end HS year	
2. Gains in growth and stature	Height, weight, skinfold thickness	W	(Beg.)/end HS year	
3. Increased vigor	Treadmill or step test	RS	(Beg.)/end HS year	Adaptation for preschool children

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
II. Perceptual-Motor, Cognitive, and Language Skills				
A. Perceptual-motor				
1. Visually guided fine-motor	CIRCUS No. 4--Copy What You See	W	(Beg.)/end HS year	Examine for culture-fairness
2. Visual recognition and discrimination	CIRCUS No. 3--Look-Alikes	W	(Beg.)/end HS year	Examine for culture-fairness
B. Cognitive				
1. Maturational indicator	Ravens Colored Progressive Matrices	W	(Beg.)/end HS year	Adaptation
2. Problem solving	CIRCUS No. 13--Thinking It Through	W	(Beg.)/end HS year	Examine for culture-fairness
3. Numerical readiness	CIRCUS No. 2--How Much and How Many	W	(Beg.)/end HS year	Examine for culture-fairness
4. Letter-number recognition	CIRCUS No. 5--Finding Letters and Numbers	W	(Beg.)/end HS year	Examine for culture-fairness
C. Language				
1. Vocabulary	CIRCUS No. 1--What Words Mean	SpS	(Beg.)/end HS year	Examine for culture-fairness
	CIRCUS No. 1a--Spanish version of What Words Mean	SpS	(Beg.)/end HS year	Development

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
II. Perceptual-Motor, Cognitive, and Language Skills (continued) C. Language (continued) 1. Vocabulary (continued)	CIRCUS No. 1b--Test of English vocabulary required in a monolingual school	SpS	(Beg.)/end HS year	Development
2. Comprehension and recall	CIRCUS No. 9--Listen to the Story	SpS	(Beg.)/end HS year	Examine for culture-fairness
	CIRCUS No. 9a--Spanish version of Listen to the Story	SpS	(Beg.)/end HS year	Development
3. Competence in use in structured situation	CIRCUS No. 10 (1a, 2)--Say and Tell	SpS	(Beg.)/end HS year	Examine for culture-fairness
	CIRCUS No. 10a (1a, 2)--Spanish version of Say and Tell	SpS	(Beg.)/end HS year	Development
	Two Person Communication Game #1 (English version)	SpS	(Beg.)/end HS year	Development
	Two Person Communication Game #2 (Spanish version)	SpS	(Beg.)/end HS year	Development
4. Competence in use in structured situation	CIRCUS No. 10 (1b, 3)--Say and Tell	SpS	(Beg.)/end HS year	Examine for culture-fairness
	CIRCUS No. 10a (1b, 3)--Spanish version of Say and Tell	SpS	(Beg.)/end HS year	Development

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
II. Perceptual-Motor, Cognitive, and Language Skills (continued)				
D. Test-taking behavior	CIRCUS No. 16--Behavior Inventory	W		Examine for culture-fairness
III. Social and Personal Development				
A. Action systems: role behaviors toward significant others and their responses				
1. Role behaviors toward peers and their responses				
a. Peer evaluation	Sociometric task: peer nominations picture naming	W, (R_{rs})	Fall K/1	Adaptation
b. Peer interaction styles	Structured observation during free play (Ogilvie and Shapiro)	W, (R_{rs})	Fall K/1	Adaptation
c. Unobtrusive peer measure	Automatic time-sampling with camera	RS	Fall K/1	Development
2. Role behaviors toward teachers and their responses				
a. Teacher-generated evaluations	Kelly role construct repertory test	W, (P_{rs})	Fall K/1	Adaptation

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
III. Social and Personal Development (continued)				
A. Action systems: role behaviors toward significant others and their responses (continued)				
2. Role behaviors toward teachers and their responses (continued)				
b. Standardized teacher evaluations	Classroom Behavior Inventory	W, (R_{rs})	Fall K/1	
c. Summary estimates by teachers	Behavior rating	W	Fall K/1	Adaptation
d. Interpretation of teachers' evaluative constructs	Large item pool for CBI (e.g., the California Child Q Set)	RS	K/1	Adaptation
	Semantic differential rating of picture stimuli	RS	K/1	Adaptation
e. Child-teacher interaction styles	Structured observation during an informal indoor task (Ogilvie and Shapiro; U.S. Commission on Civil Rights; Grotberg Appendix)	W, (R_{rs})	Fall K/1	Adaptation

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
III. Social and Personal Development (continued)				
A. Action systems: role behaviors toward significant others and their responses (continued)				
3. Role behaviors as evaluated by parents and others				
a. Parent-generated evaluations	Kelly role construct repertory test	W, (P_{rs})	Fall K/1	Adaptation
b. Summary estimates by parents	Behavior rating	W	Fall K/1	Adaptation
c. Parent involvement	Archival data	W	K/1, after years' end	Adaptation
d. Interpretation of parents' evaluative constructs	Semantic differential using picture stimuli	RS	K/1	Adaptation
e. Observers' standardized evaluations	Classroom Behavior Inventory	W, (R_{rs})	Fall K/1	
B. Characteristics of action systems: academic				
1. Child-task interaction styles				
a. Executive skills	Structured observations during individual learning tasks (Bronson)	W, (R_{rs})	Fall K/1	Adaptation

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
III. Social and Personal Development (continued)				
B. Characteristics of action systems: academic (continued)				
1. Child-task interaction styles (continued)				
b. Test-taking behavior	(See cognitive battery--CIRCUS #16)			
c. Institutional indices of success and failure				
i. Archival data on students' successes and failures	Archival data	W	K/1, after years' end	Adaptation
ii. Scales of early adjustment	Behavior ratings	W	Fall K/1 (by end of second week of class)	Adaptation
iii. Social impact	Photo naming and best/worst student nominations	W	1/2	
2. Learning styles				
a. Direction following and task completion	Mastery task (Bronson) or dual focus (Blocks')	W	K/1	Adaptation
b. Goal-setting and self-evaluation	Complex task (Crandall, Weiner)	W	K/1	Adaptation
c. Intentional-incidental learning	Modeling experiment (Portuges and Feshbach; Ross)	RS	K/1	Adaptation

See footnotes on p. 66.

00080

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
III. Social and Personal Development (continued)				
B. Characteristics of action systems: academic (continued)				
2. Learning styles (continued)				
d. Reinforcement style	Concept-switching or other learning task (Zigler)	RS	K/1	Adaptation
e. Curiosity				
i. Epistemic motivation	Unbalanced/unusual designs (Maw and Maw)	W	K/1	Adaptation
C. Action system characteristics: socioinstitutional				
1. Role-taking				
a. Perception of spatial perspective	Piagetian egocentrism-sociocentrism task	W, (P_{rs})	K/1	Adaptation
b. Perception of situational perspective	Emmerich role-pictures discrimination task	W, (P_{rs})[c]	K/1	Development
c. Perception of cultural perspective	Scott pictured value-expectation perception task	W, (P_{rs})	K/1	Development

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
III. Social and Personal Development (continued)				
C. Action system characteristics: socioinstitutional (continued)				
2. Response range				
a. Response repertory given nonpersonal stimuli (resiliency)				
i. Boundary elasticity	Concept-switching task			
ii. Multiple solutions	One of 3 tasks from the Block battery	RS	K/1	Adaptation
iii. Barrier behavior	"Stuck drawer" (Blocks')	W	K/1	
b. Response repertory given interpersonal stimuli				
i. Consequential reasoning	"What would your teacher do?" or--if not feasible--"What happens next?" (Spivak and Shure)	W	K/1	Development
ii. Conflict-resolving alternatives	The PIPS test (Spivak and Shure)	W	K/1	
D. Attitudinal constructs				
1. School attitudes				
a. Open-ended attitude expression	Interview	RS	K/1	Development

See footnotes on p. 66.

Table 2-1 (continued)

Outcome	Measure	Sample[a]	Time of Administration[b]	Measurement Needs
III. Social and Personal Development (continued)				
D. Attitudinal constructs (continued)				
1. School attitudes (continued)				
b. Structured attitude survey	Alligator game and sentence completion (PASS, Minuchin et al.)	W	K/1	Adaptation
c. Attitude toward intellectual achievement	Children's Achieve-Wishes Test (Crandall)	W	K/1	Adaptation
2. Self attitudes				
a. Open-ended attitude expression	Interview	RS	K/1	Development
b. Structured self-esteem items	Self-social Constructs Test (Ziller)	W	K/1	

[a] W = entire sample; RS = random subsample, where the subsample consists of a proportional stratified sample from the total stratified sample; SpS = special subsample, where the subsample consists of the total, or a proportional subsample of the sample for particular strata; (R_{rs}) = optional repeated measure on a random sample; (P_{rs}) = optional pretest on a random subsample of treatment children only.

[b] (Beg.)/end HS year = optional pretest, required posttest; K/1 = first year after Head Start; 1/2 = second year after Head Start.

[c] For the premeasure: only family role stimuli will be used.

Table 2-2

INDEPENDENT VARIABLES

Variables	Measure	Sample	Time of Administration	Measurement Needs
I. Variables Specifying Treatment Condition				
A. Amount of treatment, peer group ethnic composition and turnover	Questionnaire	All classrooms in sample	Late winter, HS year	Construction
B. Curriculum model	Weikart classification	All classrooms in sample	Late winter, HS year	Adaptation
C. Classroom activities and priorities of each (e.g., time in each; level of control; cultural specificity of materials)	CIRCUS No. 17: Educational Environment Questionnaire PLA-Check observational instrument	All classrooms in sample	Late winter, HS year	Adaptation
D. Natural language environment in classroom	Tizard observational instrument	SpS: subsample of English-speaking classroom	Late winter, HS year	Adaptation
II. Variables Specifying Control Condition	Questionnaire	All control children	At posttest, HS year	Minor development, construction
III. Background Variables				
A. Child	Questionnaire	W	Prior to language pretests, HS year, except for attendance record variable Attendance record, posttest, HS year	Construction

Table 2-2 (continued)

Variables	Measure	Sample	Time of Administration	Measurement Needs
III. Background Variables (continued)				
B. Family	Questionnaire	W	Prior to language pretests, HS year Check for changes, posttest, HS year	Minor development construction
C. Teacher and teacher aide	Questionnaire	Teachers and aides of all HS classrooms in sample	HS year	Construction
		Teachers and aides of all kindergarten and first grade classrooms in which treatment and control children are enrolled	Fall, K/1	
D. Center characteristics	Questionnaire	All centers in sample	Winter or spring, HS year	Minor development construction
E. Catchment area characteristics	Archival information (stratified sampling list)	All catchment areas in sample	HS year	
F. Kindergarten/First Grade Classroom and School Characteristics	Questionnaire	All classrooms, schools in which treatment and control children are enrolled	Fall, K/1	Construction

Chapter 3

HEALTH AND NUTRITION

HEALTH AND NUTRITION OUTCOMES 70

METHODOLOGICAL CONSIDERATIONS 72
  Pretests 72
  Longitudinal Evaluation 76
  Reliability of Measurement 77

PRESENCE OF DISEASE OR IMPAIRMENT 77
  Disease Vulnerability 77
  Nutritional Deficiencies 79
  Sensory or Neurological Impairment 84
    Visual Problems 84
    Hearing Problems 85
    Neurological Impairment 86

PRESENCE OF GOOD HEALTH AND NUTRITION 87
  In-Depth Evaluation 88
  Other Indicators 90

Chapter 3

HEALTH AND NUTRITION

Since its inception, Head Start has emphasized that no child shall be handicapped in his development because of a preventable or correctable health problem. Therefore, a wide variety of activities aimed at ensuring and augmenting a child's ability to exploit his environment are required of all Head Start centers. The services to be provided include diagnostic screening, hot meals, immunization, medical and dental services, psychological counseling, and health and nutrition education for parents. Good health and adequate nourishment are important goals in themselves, but in an evaluation of social competence, they gain additional importance because of the assumption that a healthy child is better equipped to learn and will be able to deal more effectively with his environment.

The conceptualization and measurement of competence in the area of health and nutrition are perhaps the most clear, concrete, and generally accepted processes of any of the outcome areas in this design of a national evaluation of Head Start. One reason for this clarity is the fairly explicit statement of the health goals and performance standards set forth by Head Start policy (OCD-Head Start Policy Manual, 1973); although there are some differences in the comprehensiveness and quality of health activities from center to center, the same basic services are provided. A second reason is the consensus among health professionals, unmatched in other outcome areas, as to the basic components of child health and the means of measuring it. Third, health variables, input and output, are often easy to quantify.

HEALTH AND NUTRITION OUTCOMES

Two outcome classes can be identified in which to measure Head Start contributions to health and nutritional status and social competence. One class is concerned with the presence of disease or impairment. A way of assessing whether children, by virtue of having participated in Head Start, have been given the potential for experiencing optimal health is to identify evidence of less than optimal health in Head Start and companion populations. Head Start's health intervention is designed to detect health problems and provide timely care or referral. If it is effective in this goal, one might expect reductions in the incidence of such childhood vulnerabilities as preventable infectious disease, stress-related disease, nutritional deficiencies, unfilled dental caries, sensory and neurological impairment, and system malfunctions.

The second outcome class is concerned with positive health, the presence of good health and nutrition. The World Health Organization currently defines health as "a state of complete mental, physical, and social well-being and not merely the absence of disease and infirmity," but this definition hardly lends itself to precise quantification. We know how to assess the absence of disease and nutritional deficiencies better than we know how to detect the presence of good health. Yet the difference in the two approaches to assessing health status is not just one of emphasis or of mirror opposites. An evaluator or mother might justifiably ask: Even though Head Start has reduced the incidence of illness among the children in the program, has the program made the children (more) healthy? "Healthy" in this sense echoes the emphasis of the World Health Organization definition of well-being and optimization of potential; it implies the ability to identify *levels* of health among those free of morbidity. Measures of such a state are not well developed or agreed upon; as Fanshel (1969, p. I3) complains, "more is known about macaroni and corsets than the health status of the population." But each person seems to have an intuitive feeling for which children are healthy, if not what constitutes a healthy child. Just as a teacher judges her children's health on such simple, though not always explicit, criteria as bright eyes and a clean nose, the evaluator grasps for similar indicators, perhaps of a more objective nature.

Part of the strategy for selecting outcomes and measures in health and nutrition is the determination of a chain of events linking Head Start health interventions with important child outcomes. There are some established cause-and-effect relationships linking specific health services to measurable effects. In such cases where a link has been established (e.g., immunizations lead to near-zero incidence of infectious disease), inputs may serve as adequate proxy measures for the outcomes they consistently produce. In other cases, the input-output link must be explored further. For instance, one member of the Rand Health and Nutrition panel, Katherine Messenger, argues that the provision of health services per se should be taken as evidence of overall better health for each child (see Appendix A). While there is a positive relationship between family income and both utilization of health services and actual health status indicators, we do not know whether a relationship between the latter two variables exists or is solely the result of the income variable link. In fact, one can argue convincingly that increased use of physicians and dentists can reflect either improved or deteriorating health status. Until the relationship between utilization and health status is clarified, utilization data remain inadequate proxy measures.

METHODOLOGICAL CONSIDERATIONS

Before going on to a discussion of specific measures, we will discuss some methodological problems inherent in the administration of the health and nutrition battery. While not unique to this battery, considerations about pretests and longitudinal evaluation become more critical here because of the greater potential reactivity of health measures.

Pretests

Consideration of the timing and frequency of the recommended health measures raises the question of the necessity of pretesting. The purposes of a pretest are twofold: assuring comparability of nonrandomly assigned experimental and control groups and measuring individual gains. The first purpose is a strong argument for a pretest if random assignment cannot be guaranteed in some centers, and the value-added design (see Chapter 7) is not used. If a treatment/non-comparable control group design is used, initial differences between the Head Start children and the control group have to be taken into account. If the value-added design is implemented, pretest data are needed from the treatment group to establish growth curves.

The second purpose of a pretest is not as relevant in the health outcome area. Unlike the other outcome areas, specific gains are not always as meaningful as the absolute level of health a child attains. For instance, a child with iron deficiency anemia may, as a result of improved nutrition, have his blood iron level raised three percentage points; but, if he is still clinically and functionally anemic, what does the "gain" signify? Therefore, for the most part, attainment of a minimally acceptable, even optimal level of health is the criterion of Head Start success rather than gains per se. Thus, if it were possible to assign children to Head Start or control conditions randomly, thereby assuring comparability of groups, a posttest design could be used as follows:

$$\begin{array}{llll} \text{Head Start} & R & X & 0_1 \\ \text{Control} & R & & 0_2 \end{array} \qquad (3.1)$$

where R represents the random assignment of subjects to Head Start and control groups and 0 represents posttests given to each group after Head Start intervention (X) or no intervention. (The left-to-right dimension of the design indicates temporal order.) While children in the control group will not have the benefit of Head Start health inputs, by virtue of their participation in the experiment, they should be assured follow-up treatment for problems detected in the posttest.

If random assignment to experimental conditions cannot be carried out at some centers, then pretesting is necessary to assess the comparability of and to correct for initial differences between the comparison groups or to collect growth curve data. Under such circumstances, problems of the contaminating effect of a pretest and of overburdening sample children with tests are introduced. First, for the Head Start sample, providing uniformly high-quality screening in the early part of the year contaminates the Head Start intervention itself. Problems may be diagnosed that might have been missed in the regular Head Start health examination, with the possible consequence that additional treatment will be provided. Similarly, screening of comparison groups is bound to affect their subsequent use of health services, since lack of referral and follow-up for diagnosed problems is medically and ethically out of the question.

The other problem with pretests in the health battery, overburdening, results from the fact that Head Start children routinely undergo health screening as part of the program. Since the screening occurs at a variety of times during the beginning of the Head Start year, allowing no standardization of testing times or intervals, and since ultimately it is the quality and success of that screening that is being judged by this evaluation, evaluation tests independent of Head Start's own procedures are mandated. However, with this independence comes a proliferation of screenings; the Head Start sample could be barraged with three different screening sessions--pretest, Head Start routine screening, and posttest.

These two considerations make a straightforward pretest-posttest design unattractive; however, these two problems can be resolved or minimized by the following design:[1]

$$\begin{array}{lcccc} \text{Head Start} & R & O_1 & X & \\ & R & & X & O_2 \\ \hline \text{Control} & R & O_3 & & \\ & R & & & O_4 \end{array} \qquad (3.2)$$

where R represents the random assignment of subjects within the Head Start and, if the design calls for one, the control group into two equal halves, one-half to receive a pretest (O_1, O_3) only and the other half to receive the posttest (O_2, O_4) only. (The row of dashes separating the comparison groups indicates that these groups are not equated by random assignment.) As in design (3.1), all control children should be assured follow-up treatment of detected problems after the pre- or posttest.

---

[1] Based on the "separate-sample pretest-posttest control group design" of Campbell and Stanley (1966).

First, in such a design, comparability of comparison groups is assessed (0_1 versus 0_3). Second, the same persons are not retested, and therefore the confounding effect of testing is avoided. Third, the number of screenings that Head Start children have to undergo is reduced. Fourth, because of random assignment *within* groups, measurements of gains or growth are possible by comparing 0_2 and 0_4, taking into account initial differences between the Head Start and control groups (0_1 versus 0_3).[1] Of course, as indicated earlier, mean differences between the two groups may not be as meaningful as a difference in proportions of each group which reach some minimum health threshold.

A disadvantage of the above design lies in the loss of half the sample to posttest analysis in order to avoid the contaminating and overburdening effect of testing. An alternative design might be the following:

$$\begin{array}{lccccc} \text{Head Start} & R & 0_1 & X & 0_2 \\ & R & & X & 0_3 \\ \hdashline \text{Control} & R & 0_4 & & 0_5 \\ & R & & & 0_6 \end{array} \qquad (3.3)$$

where one-half of each of the comparison groups would receive both pre- and posttests. *If* no contaminating effects of testing were found (0_5 versus 0_6 and 0_2 versus 0_3), then the full sample of posttests could be used, thereby increasing the strength of the inferences to be drawn. The unattractive feature of such a design is the heavy testing (three separate screenings) of half of the Head Start sample.

---

[1] The two groups are successfully equated only if unmeasured variables on which the groups might differ do not affect the measured variables and if the reliability of the measured variables is very high.

Longitudinal Evaluation

As indicated in Chapter 1, there are several general arguments against a longitudinal evaluation. In addition, problems specific to the health intervention question the appropriateness of a longitudinal design. There are at least two different assumptions about the effect of Head Start on a child's health that bear on whether a longitudinal component to the evaluation is appropriate. The first asserts that, because of the health services provided during the Head Start year, a child's health status will improve and his diseases and deficiencies will be treated and remedied; but these services end when the Head Start intervention ends, and additional improvements are therefore not to be expected. The second assumption states that the critical effect of Head Start is on the health and nutrition behavior and attitudes of parents; if permanent change in this intervening variable occurs, then continued improvement in the child's health and nutrition might be expected. This assumption presupposes a causal chain: first, that any intervention is capable of changing parents' attitudes; and second, that changes in parent attitudes ultimately affect the child's health. Neither link has been demonstrated. In fact, as far as the first is concerned, the Health and Nutrition panel affirmed the difficulty of changing well-established habits (especially nutritional habits); for, even if people wished to avail themselves of health services or change their food purchases, the economic means to do so may not be available to them. For the above reasons, a longitudinal effect of Head Start is unlikely to be found.

The higher costs incurred by repeated testing, the intrusive character of prolonged testing, and the low probability of a longitudinal effect are reasons for only one (if any) set of longitudinal measures, perhaps in the second grade. Such an assessment, however, carries with it the same problem of possible contamination from pretesting encountered previously, for now the posttest measures done at the end of the Head Start year effectively become pretest measures. This problem dictates the division of Head Start children (and control groups) into extra randomly assigned subgroups, one subgroup receiving a posttest only in the second grade.

Reliability of Measurement

The efforts of the National Health Examination Study of the National Center for Health Statistics illustrate the kinds of measurement problems that are predictable in any large-scale medical evaluation. The center's experience and strategies for dealing with reliability problems, such as standardized testing conditions, standardized training of testers, and replication of tests where error is likely to be high, are documented (NCHS, 1965; 1972b) as lessons from which researchers such as ourselves can benefit. For instance, in order to ensure uniformity and reproducibility of the data, a standardized exam environment was established since some measurements (e.g., vision and hearing screening) will vary depending on environmental influences. Instruments were checked and recalibrated periodically; two determinations of hematocrit were made per child. Physicians underwent basic training in the examination protocol, with occasional retraining sessions; in addition, consultants were brought in from time to time to observe parts of the examination to ensure that the protocol was being followed and the data were of high quality. Each of these precautions should be incorporated into the small number of the procedures in the battery in the subsample of Head Start and control children.

PRESENCE OF DISEASE OR IMPAIRMENT

Disease Vulnerability

As one member of the Health and Nutrition panel (see Appendix A) pointed out, "this is among the healthiest times in most people's lives." Young children have extremely low mortality rates and are also low in reported morbidity. For instance, the National Health Interview Survey (NCHS, 1972a) reports that children under 17 years of age suffer the lowest rate of activity limitation due to chronic conditions; the lowest rate of restricted activity from disease, injury, or impairment; and the lowest rate of hospitalization (and shortest length of hospital stay). However, acute infectious and respiratory conditions and injury occur with greatest frequency in this age group. Reduced incidence in preventable infectious disease is an outcome Head Start can be expected to affect, for immunizations are routinely given to children on entry into the program. Records of inoculations for rubella, rubeola, and mumps, obtainable with little effort, may be used as outcome measures; for, as a result of active immunization, a child possesses an effective level of protection against such preventable diseases in keeping with his age.

Equally easy to gather, but notoriously inaccurate and indeed unnecessary if inoculations are accepted as a proxy measure, are data on the actual occurrence of these common illnesses against which the inoculations are supposed to protect the child. Since the incidences of these diseases are documented for various income and ethnic groups by the National Center for Health Statistics and the Center for Disease Control, incidence levels can serve as controls. Other infectious diseases (diphtheria, pertussis, tetanus, poliomyelitis) do not occur with sufficient frequency to make recording their presence in Head Start children useful as a measure. By the same token, the incidence of tuberculosis, except in special populations (e.g., Native Americans, migrants), and incidences of major system diseases are also too low to expect observable differences. For the above considerations, *the input measure of numbers of inoculations for rubella, rubeola, and mumps performed at Head Start centers as opposed to numbers among control children is recommended. For the oversampled population of Native Americans* (see Chapter 7), *the number of positive tuberculin skin tests should be an additional measure of disease incidence.* Head Start would be judged successful in this outcome area if more Head Start children than control children were effectively immunized and if fewer Native Americans in the program than outside it were detected as positives and fewer previously known positives in the program had gone untreated.

Another class of health problems to which preschool children may be susceptible is stress-related disease. The logic supporting a possible Head Start effect is that by providing the child with a healthful environment, by making him more content with himself and more competent, and by offering social counseling to the family, the program may reduce stress at home and thereby its consequences for the child. An obvious candidate in this disease category is asthma; while it is now believed that asthma is mainly a direct result of allergy, it has been suggested that the condition's onset can be triggered by emotional stress. However, panelists felt it was unlikely that the Head Start intervention in the social welfare of the family would be sufficiently powerful to counteract current family relationships and to reduce the incidence of asthma.

Reduction in the number of child abuse incidents--also related to family stress--is regarded as a fruitful measure of Head Start effectiveness, but it is difficult to capture. While medical histories or emergency room records might provide information on the occurrence of abuse, these measures have obvious reporting difficulties since parents or other caretakers are unlikely to reveal the causes of injuries. Similarly, reduction in accidents, while considered to be an important outcome, suffers from the lack of satisfactory available measures. First, there is a problem of definition as to what constitutes an accident to be recorded for the purposes of the evaluation. Second, there is the problem of how to obtain the number of occurrences of the incidents to be recorded. Again, some reliance would have to be placed on parent recollection, an unsatisfactory procedure. Third, there are problems of interpretation of the data obtained from, for example, emergency room records or parent reports. Is a child whose parents report more injuries or take him to the emergency room for injuries more often less healthy than one whose parents neglect him and his injuries?

Nutritional Deficiencies

In the outcome area of nutrition, less than optimal status is measured in terms of deficiencies in need of remediation rather than in terms of diseases. One primary focus of the Head Start health program is the child who has not been well fed and is therefore malnourished. Early and severe malnutrition, as evidenced in extremely impoverished populations in underdeveloped countries, has been linked to retarded cognitive development (National Academy of Sciences, 1973). However, the effects of marginal malnutrition have not been demonstrated. Nonetheless, it has been suggested that poor nourishment may, by influencing attention, play an important role in intellectual test performance. Therefore, nutritious snacks and hot lunches are mandated by Head Start performance standards; and nutritional problems of center children and their community are taken into consideration in planning the nutrition program. As clearly pointed out in a position paper of the Food and Nutrition Board of the National Academy of Sciences (1973), the influence of hunger on school performance or, conversely, the efficacy of breakfast, snack, or lunch programs in improving school performance has not been documented. Careful studies on both the physiological and psychological effects of hunger are needed.

The methods used in the assessment of nutritional status will vary in keeping with the pretreatment level of nutritional health expected in the populations under study, their socioeconomic characteristics, and expectations or nutritional goals to be projected for the population. In any population of preschool children in the United States, the nutritional needs should be optimally met, and thus the highest level of nutritional health is an appropriate and defensible goal.

In order to differentiate between malnutrition caused by faulty diet and malnutrition caused by metabolic or other medical factors preventing absorption or assimilation of food nutrients, a diet history is desirable. Such a history also provides information on child and family eating habits that might be addressed by the nutritional education component of Head Start. A history may be achieved in several ways, none of which is very satisfactory. The first way is a 24-hour recall in which the child or his parents sits down with a nutritionist and lists the foods and quantity of foods the child has eaten in a 24-hour period. A second method is that the parent records everything the child eats over a three-day period and then submits the list to the nutritionist for evaluation. While there are even computer programs available by which this dietary information can be analyzed, it is generally believed that records are not very reliable and that there is a strong tendency for parents to make the child's diet sound better on such measures than it actually is.

A third method, probably the most exact but also highly intrusive and expensive, involves placing a competent nutritionist in the home of the child for a definite length of time so that the food consumed by the child during this period might be recorded. However, the difficulty in this situation is that the presence of a visitor may result in a temporary improvement of perhaps formerly inadequate meals and therefore confound any real effects from Head Start. Thus, although several panelists attested to the importance of dietary intake as an outcome, assessment of changes in eating habits are not likely to be feasible even on small subsamples because of acknowledged measurement complexities.

Rather than judge the adequacy of a child's diet by measuring the nutrient input, one could measure the presence of adequate diet through the absence of nutritional deficiencies, such as vitamin or iron deficiencies. Laboratory tests for these indicators of malnutrition are common and can be incorporated easily into any screening protocol.

Iron is the most commonly limited nutrient in the North American diet. Iron deficiency is widespread throughout many sectors of American society because of the physiological limits of assimilation of iron from the diet and restricted dietary intakes. Shown to be the principal cause of nutritional anemia in the United States (Healy, 1973), it is known to occur with greater frequency in disadvantaged (particularly black) children, rather than middle-class children (Center for Disease Control, 1972b). Because cases of severe chronic anemia with its well-known complications are rare in the United States, this evaluation is primarily concerned with the effects of mild and moderate iron deficiency on child development. In studies conducted in Welsh mining communities, Elwood and Waters (1969) found no clear relationship between somewhat low levels of iron in the blood and clinical signs of dysfunction. However, some recent studies have shown significant effects of such mild or moderate deficiency levels on attentiveness and learning ability in children (Howell, 1971). Further, when the deficiency is corrected, improvements are noted in both outcomes (Sulzer, 1971). In fact, there is preliminary evidence that nutritional and nutrition education intervention can have an effect on the hemoglobin concentrations of a Head Start population (Healy, 1973). Reasons for the presence of iron deficiency in deprived preschool populations can range from poor family buying habits or lack of knowledge about proper diets to insufficient money for purchasing proper foods. The former reasons are certainly addressed in the Head Start philosophy. Thus, *measures of iron deficiency meet the selection criteria of prevalence, relevance to Head Start activities, and importance.*

Hemoglobin determinations are traditionally used in the assessment of anemia, but hematocrit data (volume of red blood cells as a percentage of whole blood) correlate well with hemoglobin data (r = 0.85 in the Ten-State Nutrition Survey, Center for Disease Control, 1972b) and can be used in their place. Both measures are reliable, inexpensive, and readily performed in any laboratory. In fact, a new measure, serum ferritin, represents an important advance in evaluating iron nutrition, reflecting total body iron stores with greater specificity. Whatever the test, the wide range of normal values and the disputed functional significance of abnormal values make interpretation imprecise (Elwood and Waters, 1969; Kessner and Kalk, 1973); an acceptable value for one person may indicate deficiency for another. For the purposes of the present evaluation, the criteria for iron deficiency anemia should be at least a 33 percent hematocrit and a low serum ferritin; success of the Head Start program may be judged by the lower incidence of anemia among participants.

With the exception of iron, the mean dietary intakes of infants and young children have been shown to be sufficient to meet nutrient standards (Center for Disease Control, 1972b). Vitamin A deficiencies have been found in some studies, but it is generally agreed that no functional effects have been associated with these findings. Measures of vitamin C deficiency seem to indicate only very recent intake and are therefore not reliable measures of general diet adequacy. Protein deficiency, while not likely to be prevalent in the Head Start population as a whole, may be indicated for extremely low-income children and Native American children. Along with TB screening, *serum albumin determinations should be conducted for the Native American subpopulation.* The incidence of lipoprotein imbalance, measured by cholesterol level, may be considerable (either through genetic predisposition or because of high fat intake); however, since Head Start is not addressing this problem, it is not likely to be able to affect cholesterol levels.

Dental caries, while not deficiences, are another aspect of nutritional status and diet adequacy. Both dental caries and periodontal disease are related to the intake of refined carbohydrates (e.g., pastries, candies, soft drinks); hence dietary habits have a major influence on dental health. The Ten-State Nutritional Survey (Center for Disease Control, 1972a) data also suggest that poor dental health is associated with poor levels of care. *The Head Start program, through its educational and dental services components, addresses both of these sources of poor dental health; the dental examination is therefore recommended for inclusion as a measure in this evaluation.* The criteria of Head Start success on this measure might be fewer untreated caries among Head Start children, a clear indicator of improved care; and fewer total caries, treated and untreated, an indicator of improved diet and oral hygiene.

While measuring and noting changes in the negative aspects of dental health (e.g., decayed, missing, and filled teeth), a dental examination also indicates the degree of the child's improvement in good health habits. Thus, dental health can be viewed as a positive as well as negative health outcome; improved dental health reflects improved nutrition, improved professional care, and improved personal care.

The dental examination will need a standard protocol to be followed for all children (e.g., McElroy and Malone, 1969), and participating dentists will need to be trained in the procedures of the examination to ensure uniformity. As with health screenings, Head Start centers perform dental examinations as part of the health program. But since these examinations are not given uniformly on entry but throughout the year, and they are the treatment, they cannot appropriately be used as either a pretest or a posttest.

Iron deficiency and dental health appear worth pursuing in a national evaluation. The ease of administration of the hematocrit screening and dental examination commend the measures for the basic sample battery.

Sensory or Neurological Impairments

An obvious prerequisite of adequate functioning in a school or any other environment is intact sensory systems, particularly vision and hearing, so that the child can communicate effectively with others and thereby learn social and academic skills. Neurological damage (e.g., minimal brain damage), if undiagnosed, can result in unrealistic parent and school expectations and concomitant emotional stress for the child. Thus, screening and remediation (where possible) of these impairments are major services of the Head Start program, the benefits of which are long-lasting for the affected children.

Visual Problems. Visual disorders occur with unknown frequency; it has been estimated that 10 to 13 million school children in the United States require some professional visual care (NCHS, 1970). The most prevalent visual disorders are refractive errors: hyperopia (farsightedness), myopia (nearsightedness), astigmatism, and anisometropia (unequal refractive power in the two eyes). The majority of pediatric visual problems can be detected by any of the commonly used screening techniques of visual acuity based on the Snellen method. Such techniques are *not* diagnostic but do indicate the need for further evaluation. Published studies show that 5 to 10 percent of preschoolers are correctly referred for care ("true positive"); the comparable rate for five-year-olds is approximately 15 percent, with an increase of 1.6 percent annually in elementary school (Kessner and Kalk, 1973). About 15 percent of the children between 3 and 16 years of age wear corrective lenses (NCHS, 1969). The prevalence of uncorrected visual problems among children does not seem to vary with socioeconomic status, but severity of visual defects does show an inverse relation to family income (Kessner and Kalk, 1973).

Functionally, peripheral eye disorders such as those indicated by Snellen measurements play a disputed role in learning and reading disabilities--e.g., dyslexia. While some maintain their effect on performance in these areas, a recent report of the American Academy of Pediatrics does not support a causal relationship; normal achievers suffer the same incidence of peripheral eye disorders as children with learning disabilities (Kessner and Kalk, 1973). However, although perhaps not the principal cause of reading failure, correctable visual disorders obviously contribute. If corrected at the earliest possible time, adequate visual acuity, a necessary though not sufficient condition for learning, can be ensured.

For the present evaluation, *it is recommended that all children in the basic sample be screened for monocular acuity, with and without corrective lenses, with the Snellen test.*[1] Without-lenses screening indicates the base prevalence of deficient acuity in the Head Start and control samples included in the evaluation; with-lenses screening reveals both the proportion of problems that have been treated and the efficacy of the correction. The criterion of Head Start success will be a higher proportion of corrected problems among Head Start children.

Hearing Problems. For decades educational professionals and psychologists have assessed the effects of hearing loss (specifically loss due to middle-ear infection) on academic performance. Hearing-impaired children were more often found to be retarded in reading, spelling, and language ability (Polling; 1953; Schonell and Schonell, 1946; Young and McConnell, 1957). Significantly, failure on an audiometric test, whether the condition detected was caused by severe permanent hearing loss or mild temporary loss, seems to be a good predictor of below-normal school achievement. Concomitant effects contributing to poor achievement are limitations in the child's ability to communicate with others and in his personality development (Fisher, 1966; Kessner and Kalk, 1973). For instance, a child's ability to place labels on objects influences his ability to recall objects and to organize stimuli (Leeper, 1935; Vernon, 1955). This relationship suggests the importance of auditory cues to cognitive development. Further, documentation of the social and emotional problems of deaf children indicates the importance of hearing to that area of development (Kidd and Kidd, 1966). Therefore, detection of auditory handicaps with the opportunity for improvement in communication skills is a valuable contribution of Head Start's health program to the overall social competence of the child. The program's effectiveness can be straightforwardly measured in numbers of hearing problems not detected or referred for treatment among Head Start versus control children.

---

[1] For ages 4 to 5, the Snellen Single Target "E" cards (at 20 feet) are typically used; for ages 6 to 11, the Linear Snellen "E" chart (at 20 feet) is used.

For the present evaluation, *the traditional screening technique, the pure-tone audiometric screening test, is recommended for the basic battery*. It should have a reference level for failure of 15 decibels (at frequencies of 250, 500, 1000, 2000, and 4000 Hz) and a latitude of five decibels to allow for imperfect environmental conditions (i.e., other than a soundproof room). Use of experienced screeners is recommended to assure the recording of only true responses For instance, variable intervals for presentation of the tones is a mandatory procedure. Before each test the screener should examine each of the child's ears for for foreign objects (e.g., wax, marbles, sand) that may hinder performance on the test; if anything is found in the ears, the child should be referred to a physician for cleaning before the screening is conducted. For this reason, nurses trained in audiometric screening may make the best screeners.

Securing and maintaining a young child's interest is a major problem in measuring auditory acuity. Standard audiometric tests usually show increasing acuity with age among children (Kidd and Kidd, 1966). Because of the possible confounding of results by lack of attention, however, it is unclear whether this trend indicates real auditory acuity improvement or only improvement in the child's cooperation and ability to follow directions. Therefore, the audiometric screening session should be made as interesting as possible to the child without turning it into too much of a game.[1] For children under five years of age, a more elaborate response than raising their hand when they hear the tone may be appropriate. For instance, moving beads on a string or dropping marbles in a can may hold a young child's interest longer.

Neurological Impairment. Aside from primary sensory impairment, remediated by glasses or hearing aids, there are neurological deficits that can deprive a child of his full potential for learning. For instance, the ability to pay attention is a neurological milestone in a child's development. Tasks such as reaction time to visual and auditory stimulation, tracking, digit-span repetition, and tests of habituation and distraction assess this ability. Another signal of neurological development is hemispheric specialization. The demonstration more than a 100 years ago that language functions are almost always located in the left hemisphere of man has led to innumerable studies that further delineate the functions of each hemisphere and the interhemispheric cooperation essential in certain coordinations. Attention to this problem in developmental studies has generally focused on the emergence of laterality (usually right-handedness) by some stipulated age and on the problem of "mixed dominance," although the question of precisely what the dominant eye does has not been adequately studied. It is apparent now that the preferred hand is the one that wields tools--pencils, pens, forks, etc.--but that it is the concomitant development of the nondominant (usually left) hand that may be indicative of some lag in development or other pathology. Adaptations of simple neurological tests--e.g., repetitive finger tapping, serial opposition of finger to thumb, and repetition of digits backward--may indicate whether socioeconomic deprivation is associated with laterality differences or other symptoms characteristic of slowed neurologic development. Unfortunately, the available norms for many of these tests do not differentiate children by socioeconomic levels. Furthermore, it is unclear whether any potential problems indicated by the test results can be overcome by intervention in a nonclinical setting. Therefore, *we suggest that Head Start's success in identifying, not remediating, such deficiencies is the appropriate measure if any such measure is included in the present evaluation.* No separate neurological measures need to be included in the test battery if a physical examination including tests for neurological impairments is performed (see below).

---

[1] Although use of a warble tone would make for a more interesting sound presentation, studies have indicated the success of pure-tone testing for populations four years and older. Furthermore, the warble-tone test has not been standardized.

PRESENCE OF GOOD HEALTH AND NUTRITION

The other approach to assessing health status and thereby the effectiveness of Head Start's health efforts is to identify indicators of good health rather than the absence of specific illnesses, deficiencies, or impairments. Such measures attempt to gauge status or change in status that reflect positive improvement, positive health. However, there are fewer reliable indicators of good as opposed to less than optimal health. As pointed out earlier, the state of the art for measuring positive aspects of health is less advanced, lacking objective, quantifiable indicators.

In-Depth Evaluation

One such measure can be based on a complete physical examination and medical history. Besides assessing the presence of idiosyncratic deficiencies or impairments that would go undetected without an in-depth assessment of health status, the physical examination also allows the physician to see the whole child, his infirmities and strengths. For instance, potentially the examination can note such positive features as good stature, strong heart beat, clear complexion, good muscle tone, and obvious energy as well as system dysfunction. However, the physical history and examination as a measure raises several problems. First, the cost in time and manpower (approximately one hour and $10 per child per exam) makes it problematical for the full sample of children. However, a subsample of children could reasonably be examined in such a manner.

Second, there is a need for comparable and aggregatable information on each child in the evaluation. Itemized diagnostic procedures to be covered could be adopted from which the physician would deviate only to follow up important clues. While no standardized protocols have been developed (or at least promulgated by the American Academy of Pediatrics), pediatric textbooks (e.g., Silver, Kempe, and Bruyn, 1973) offer such a procedural account. But standardized procedures alone do not assure aggregatable or reliable data. A physical exam is, by its very nature, an intuitive, individualized search by the physician for problems and for an overall indication of the child's health. A method of aggregating from this intuitive data collection should capitalize on the physician's ability to integrate and interpret history, examination, and test data. The outcome measurements in such a potentially data-rich exam could be judgments of system intactness and health, expressed in a summary index. Several past attempts to define a health index have emphasized measures of functional adequacy rather than disease, the ability to fulfill the requirements of a social role appropriate to age and sex (Sanders, 1964; WHO, 1957; Fanshel, 1969; Sullivan, 1966; Mahoney and Barthel, 1965; Rosnow and Breslau, 1966; Sokolow and Taylor; 1967). For instance, as a gross index of the child's functioning, a five-point scale could be developed indicating, for the variety of systems explored in an exam, the degree to which the child has attained a level of optional functioning in that system. System scores could be aggregated into an overall index of health with weight being assigned based on the system's relative importance to the well-being of the child. While the use of a summary index is an interesting possibility, its feasibility for the present evaluation is questionable. The reliability of such subjective indexes among physicians is likely to be low unless the same physician performs all exams. Further, if all systems are included in the index, the score is liable to be swamped by systems upon whose functioning Head Start can be expected to have little or no effect (e.g., diabetes). The measures recommended thus far already reflect those systems for which a Head Start effect is expected (e.g., eyes, ears).

In view of Head Start's role as a diagnostic and referral agent, a more reasonable measure of its effect on health status might be: (1) What proportion of Head Start children compared with control children have been examined by a physician during the Head Start year? and (2) What proportion of health problems (nonacute) have been diagnosed properly and referred for treatment? Measure (1) is a straightforward monitoring of one of Head Start's guidelines, to provide physical exams to every child in the program. No control group measure is required if current data from the National Center for Health Statistics can be obtained on the annual utilization rate of physicians by children between 4 and 5 years of age from low-income families. If we assume that seeing a physician is sufficient evidence of better health or better health care, then measure (1) can tell us whether Head Start children are better off than control children. Stronger evidence, however, would be provided by the addition of measure (2). If the quality of the Head Start exam is adequate to diagnose most or all health problems and referrals for treatment are made, Head Start has carried out its mandate and presumably the child's health care is under better management than if the problems were not diagnosed and follow-up treatment were not suggested. The problem with measure (2) is that it is feasible for the Head Start group only. The medical records required for adequate and reliable (more reliable than mothers' verbal reports) comparison of Head Start and control children are not dependably accessible for the control group. However, if within the Head Start group not only more exams [measure (1)] but high quality exams [measure (2)] are being provided, then we can assume Head Start children are receiving better care than non-Head Start children. If, however, the quality is not high, as indicated by measure (2), then we are unable to make any meaningful comparison of Head Start and control children.

The contribution of a physical exam to the battery of health measures is limited. The exam is recommended as an optional measure for quality monitoring on a subsample of Head Start children.

Other Indicators

Growth and stature (i.e., height and weight) are considered by many to be good overall indicators of good health and nutrition. The persistence of height and weight differences between persons of higher and lower income levels from childhood into adulthood indicates the cumulative effects of poverty on growth and maturation (Center for Disease Control, 1972a). Although part of the optional complete physical examination for a subsample, the measures can be applied to the entire basic sample.

Malnutrition is most often judged by measurements of height and weight for age, for which standards have been established for children in various age groups. Suggestive evidence, in fact, exists that different standards should be developed for children of different races--e.g., black children are taller than white children (Center for Disease Control, 1972a). However, in the proposed evaluation, where a randomly assigned control group design or a value-added design will be used, Head Start children can be compared with racially and socioeconomically comparable children not exposed to Head Start benefits. In either case, norms developed for a Head Start population would be irrelevant to interpreting relative health and nutritional status.

Two factors must be recognized as compromising the current usefulness of height and weight data. First, the accuracy of measurements is often open to question because of inadequate training of personnel and the use of poor equipment. Second, height and weight are gross indicators of health and nutrition and therefore are most sensitive to extreme cases of malnutrition.

Measuring skin-fat fold thickness at selected body sites (triceps, subscapular) provides a clinical index of body fat. The importance of this measure lies in its more accurate assessment of general body build.

Although children from higher-income families tend to have slightly greater head circumferences, retardation of growth in head size results only from marked degrees of nutritional deprivation extending over prolonged periods. Therefore, such a measure is not seen as a useful one for the present evaluation.

Physical, nutritional, and mental well-being is also manifested in good muscle tone and maintenance of healthy cardiovascular function. *A treadmill test or steptest measure of vigor through rate of recovery should be developed.* According to one panelists, Dr. James Carter of Meharry Medical College, these tests have been successfully used on an experimental basis with adults and older children. Adaptation of the tests to preschool children is needed so that they can adequately differentiate children on level of vigor. The vigor or fatigue that would be thus measured is purely physical; linkages to attention span or cognitive fatigue are not warranted, given present lack of knowledge on any correlations.

Use of the services of a physician or dentist is often suggested as a measure of good health. As indicated earlier, it is highly related to family income; the higher the income, the more frequent the visits per person per year (NCHS, 1968). There is also a stong association between income and health status indices, thereby creating a possibly spurious relationship between health care utilization and health status. Furthermore, utilization data place the emphasis on input characteristics of the health system, not on its end product, better health. In fact, the link between input and output in health is not a clear one. While many will argue for the implicit benefits of regular visits to the physician, recent studies have shown no relationship between quality of care provided and health status as the outcome (Kessner, Snow, and Singer, 1974) or even a relationship between whether the condition was treated and the outcome (Brook, 1972). The fact that Head Start offers preliminary medical and dental services and referral to appropriate outside care when needed may be a benefit in itself; but unless utilization of these services results in a better showing on the health status indicators recommended above, its measurement alone is not useful. For these reasons, *we do not recommend number of visits to a physician or dentist as a measure to be included in the basic battery*, either during the Head Start year or as a follow-up.

In summary, the following measures are recommended for inclusion in the health battery:

- Records of immunization (rubella, rubeola, mumps).
- Incidence of TB among Native American subpopulation.
- Hematocrit and serum ferritin.
- Serum albumin determinations for Native American subpopulation.
- Dental examination (for treated and untreated caries).
- Snellen test for visual acuity.
- Pure-tone audiometric screening test.
- Optional: Complete history and physical examination for a subsample of Head Start children.
- Growth and stature (height, weight, skinfold thickness).
- A vigor measure.

Chapter 4

PERCEPTUAL-MOTOR, COGNITIVE, AND LANGUAGE OUTCOMES AND MEASURES

CRITERIA FOR SELECTION OF PERFORMANCE DIMENSIONS 94
  Measurement of Skills 94
  Maturational Indicators 99
  Criteria for Selecting Tests 101

CIRCUS INSTRUMENTS AND THE NATIONAL SAMPLE 106
  Focused Research Studies 109
  Preferred Instruments 111

PERCEPTUAL-MOTOR SKILLS 112

COGNITIVE SKILLS 115
  A Maturational Indicator 116
  Visual Recognition and Discrimination 117
  Letter and Number Recognition 119
  Problem-Solving Skills 121
  Quantitative Concepts 124
  Metacognitive Competence 127

LANGUAGE SKILLS 128
  Vocabulary Development and Knowledge 132
  Comprehension and Recall of Oral Language 133
  Competence in Language Use: Descriptive and Narrative Use .. 135
  Competence in Language Use: Functional Use and Use in
    Unstructured Situation 140
  Metalinguistic Competence 142
  Assessing Test-Taking Behavior 143

INSTRUMENT SELECTION AND EVALUATION DESIGN 145
  Pretests 146
    Practice Effects 146
    Intervention Effects 147
    Costs 147
  Longitudinal Study 148

Chapter 4

PERCEPTUAL-MOTOR, COGNITIVE, AND LANGUAGE OUTCOMES AND MEASURES

This chapter offers a rationale for the selection of Head Start outcome measures in the domains of perceptual-motor, cognitive, and language development. The first sections of the chapter explore the relevance of measurement in these three domains to the measurement of social competence in the proposed evaluation. They also deal with the criteria that predominated in our decision to focus on certain performance dimensions and the criteria determining selection of preferred instruments. Subsequent sections of the chapter describe specific instruments and why they were selected. Some are intended for the entire national sample; also mentioned are some that seem appropriate for subsamples or focused separate studies.

The final section of the chapter addresses some issues of the relationship between test selection and evaluation design. Does the recommended battery imply a preferred strategy for subsequent data analysis? Are all tests to be administered at both pretest and posttest, or will some be given only at pretest or only at posttest? And will test selection for the national sample be contingent in part on continuity with later follow-up testing in the schools, exploring longitudinal stability of gains? Some of the answers to these questions are still uncertain, resulting in options for the OCD and the primary evaluation contractor in the coming months. But where answers can be made explicit, or where they have been implicit in the process of selecting a basic battery, these answers are presented and discussed.

CRITERIA FOR SELECTION OF PERFORMANCE DIMENSIONS

Measurement of Skills

At least some of the effect of any good Head Start program is reflected in what the OCD *Policy Manual* terms "the enhancement of the child's mental processes and skills with particular attention to conceptual and verbal skills."[1] Such skills contribute jointly to our impression of the child's social competence, constituting what for many are the most important dimensions of program-related growth. Traditionally, Head Start evaluations have placed heavy emphasis on the measurement of mental skill development, in large part because this aspect of development is closely related to subsequent performance in school. For a variety of reasons that have been well summarized by Anderson and Messick (1974, pp. 287-288), however, accurate measurement of social competence is difficult even in a clearly defined area such as school readiness. Among the problems are these:

1. Distinguishing between behaviors that are prized by many segments of the society across a large number of situations and behaviors that are not necessarily universally admired or are differentially appropriate to different situations...;
2. Distinguishing between proficiency and performance and between maximal and typical performance...;
3. Recognizing that variables may have different meanings--and thus different implications for social-educational action--at different levels of intensity or in their positive and negative ranges...;
4. Distinguishing between the positive components of social competency (the characteristics we can agree we want the child to have or develop) and negative characteristics which may serve as obstacles to learning, development, and societal adjustment...;
5. Identifying different classes of variables in terms of their developmental trends...;
6. Recognizing the importance of defining and assessing social competency in dynamic as opposed to static terms...;
7. Making explicit the relationships between program goals for parents and program goals for children.

---

[1] *OCD-Head Start Policy Manual*, January 1973, p. 7.

These issues are enduring, and all of them must be considered in the process of Head Start instrument selection. In addition, it is important to be clear about the conceptual framework within which child performance is measured. Although the three domains considered in this chapter are three quite separable major research areas in child development, there is substantial overlap in their contribution to social competence. Perceptual-motor, cognitive, and language development are closely related as reflected in children's Head Start performance and as tapped by various preschool assessment instruments. Most skills, especially those assessed by individually administered tests requiring the child to perform some task, involve a complex interaction of all three domains of mental process. A measure designed to assess classification skills, for example, usually also requires some level of linguistic proficiency on the part of the child and some rudimentary perceptual-motor capabilities. Most preschool tests of mental competence turn out to be tests not of specific component processes but of a wide range of capacities and behaviors.

One useful way to think about measures in the domains of perceptual-motor, cognitive, and language development is according to whether the skills they assess tend to be maturation-bound or teachable. Imagine a continuum of competencies ranging from the most maturation-related to the most amenable to preschool influence. It is clear that some aspects of mental development, which for purposes of this chapter will be called maturational indicators, are capabilities that cannot readily be taught, and in normal children they will emerge regardless of interventions. These emergent processes, even if they can be speeded somewhat in their onset, are aspects of mental process grounded primarily in physiological development. In this category, for instance, is latency speed for various perceptual-motor tasks. At the other end of the continuum are a group of skills that clearly can be taught and tend to develop only if the child is exposed to an intervention teaching them.

In the domains of perceptual-motor, cognitive, and language development, the intricate problem that faces us when measuring Head Start effects is that the skills relating most closely to the child's role as learner are not neatly at one end or the other of this imaginary spectrum.

Instead, they tend to fall somewhere toward the middle, requiring various maturation-related capacities but also being fostered by specific teaching. In fact, any model allocating maturational and program-related components of skill attainment would have to be far more complex than has been suggested; there are myriad ways that specific maturational sequences could interact with specific pedagogical treatments to bring about skill attainment across the interval of the Head Start year.

The teaching of basic letter-decoding skills provides a good example. Clearly one aspect of decoding is maturation-related, having to do with age-related cerebral functioning. A child below a certain age (although younger in some cases than many may think), cannot be taught to match phonemes with their written equivalents. After a later age, however, most children can easily be taught this skill, and many learn it on their own. Explicit teaching of decoding to children who are capable of learning the skill and are on the threshold of attaining it has been a major curricular emphasis of reading readiness programs over the years. It remains "teachable" in the sense that it is a skill to be mastered with systematic training after the age of capacity for attainment has been reached.

It is important to be able to group Head Start outcome measures in the child performance realm according to whether those measures access skills that are teachable in the sense that the child is ready for a pedagogical input. The program can be expected to affect the attainment of only those skills with a large intervention-related component. It would not make sense, for instance, to assess a program for four-year-olds according to how well it enabled them to attain concrete operational thinking, which cannot be expected to evolve until the age of six or later. Because Head Start can be held accountable only for those aspects of child development on which it can exert some leverage, measures in the perceptual-motor, cognitive, and language development domains must be closely related to teachable aspects of social competence. This means that some of the dimensions of mental capacity normally explored in IQ measurement and other stable trait measurement are given fairly low priority. Rather, there is an emphasis on the assessment of "leverageable" and behavioral areas of program effects. Fortunately, many social competence skills in the areas of perceptual-motor, cognitive, and linguistic process *are* teachable, or largely teachable, in the preschool.

Preference for dimensions of mental growth that are both malleable and closely related to behavioral and role-defined aspects of social competence is reinforced by the knowledge that those tests to be given to the entire national sample must be sensitive to aspects of program gain that *all* Head Start children, regardless of program type, can reasonably be expected to learn. If the proposed evaluation were one in which there were highly specialized programs, emphasizing specific curricula and teaching methods for specific areas of mental growth--principally a comparison of structured treatments--it would be reasonable to expect gains or shifts in highly selective skill areas. If one program offered a curriculum designed explicitly to teach digit-span memory and another did not, for instance, then it would make sense to give a test at the end of the Head Start year assessing gains in this area. In the absence of such a highly focused curriculum, however, it is unlikely that such specialized program-related changes can be expected. The evaluation we propose spreads a wide net, with the same basic inventory of tests administered to all children in the national sample regardless of program type. Information on program variables should be collected, however, to investigate differential effects; it is well known that in their general characteristics, some of which may be important to outcomes, Head Start centers can vary greatly from each other in the field. In addition, selective in-depth study of outcomes is proposed for subsamples and focused research studies.

Even the most face-valid program effects must be carefully interpreted. Anderson (1973, p. 10) has pointed out some of the dangers in a simplistic analysis of gains:

> It is easy to go along with the notion that knowing more letters (or colors or numbers) is better than knowing fewer, and it makes sense to give the child who knows more letters a higher score. However, such variables as response latency are not so easily interpreted; while quick responses may indicate lack of reflection, very slow ones may be more indicative of obsessiveness or fatigue than of reflectivity.

> We know, too, that some dimensions may be bipolar, and extreme behavior at *either* end may be maladaptive. (The attempts to assess "self-concept" have suffered from failure to take account of such possibilities.) The fact that different variables may show different developmental trends is relevant here, too. For example, some abilities may increase with age and training (perhaps tapering off at later ages or with lack of practice), while others may decrease with maturity, or be cyclical, or remain fairly constant across wide age spans.

Maturational Indicators

The preference for measuring strictly behavioral and program-related dimensions of growth requires one important counterbalance. Another purpose for measurement in the perceptual-motor, cognitive, and language domains, especially because measurement here is more apt to be reliable than in the socioemotional area, is to gather a minimum amount of baseline data that is expressly maturation-related, to be used as a covariate control or moderator variable in the analysis of Head Start gains. Along with the major strategy of measuring practical and leverageable dimensions of child performance, then, a minor or secondary strategy should be adopted of collecting a minimal sufficient set of data on maturational indicators or developmental markers. Two kinds of data are needed; first, those that give some indication of a child's *level of development*, and second, those that can give some baseline estimate of at least one or two stable dimensions of *ability*, since preschool children vary greatly on both. There are severe constraints on obtaining direct measures of innate mental capability. As Mercer (1974, pp. 21-33) states:

> An individual's genetic potential is always expressed through behavior acquired in a social and cultural setting, his phenotype. Thus, all tests are basically measures of achievement and all test scores are influenced by a wide variety of environmental factors as well as the person's innate capacity for learning.... [Hence, we agree with] the fundamental premise that all tests are achievement tests which can be interpreted as measures of aptitude only when an individual's performance is being compared with others who (1) have had similar opportunities to learn the skills and information covered in the test, and (2) have been similarly motivated and rewarded for learning those skills.

One approach might be to select clinical assessment measures designed to identify high-risk children in the Head Start population according to diagnostic techniques originally devised for use in screening for learning disabilities or mental or behavioral defects (following the pathology model). This approach would result in a particular kind of moderator data--essentially categorical--isolating children who for one reason or another were in trouble developmentally from children who were not. The other approach (the norm-based model) would use tests developed originally as norm-referenced instruments but using pluralistic norms. This would enable children to be ranked on a continuous distribution within their appropriate reference groups according to the trait measured.

We have received much thoughtful comment about the possibility of including diagnostic and clinical measures in the cognitive effects battery to screen for particular learning problems. But the purpose of the proposed Head Start evaluation does not quite fit under this rubric. The evaluation does not presuppose a deficit model of cognition or assume that Head Start children are clinically "at risk." Instead, the evaluation is aimed at understanding what Head Start is accomplishing with a group of healthy and normal children.[1] Sponsors also want to understand something about what program components cause what kinds of effects for what kinds of children, so that more effective programs might be designed in the future. In this regard the goals of Head Start are really too complex (and even at times contradictory) to be fitted into a simple model of remedial intervention aimed at diagnosed deficiencies.

We recommend that the measurement battery in the perceptual-motor, cognitive, and language domains not use measures whose principal intent is to assess learning disabilities or deficiencies, or tests designed for use in clinical assessment. Instead, we recommend that the issue of the child's biological intactness be assessed only through the health and nutrition measures suggested in Chapter 3 and through a single cognitive instrument to be used as a maturational indicator or baseline moderator variable of the second, norm-based type. Otherwise, the child's intactness is assumed, and the issue of "remediation" is rejected as an inappropriate conceptual framework within which to interpret Head Start effects.

---

[1] The issues involved in assessing Head Start benefits for the 10 percent of handicapped children included in the program are quite separate. They are discussed in Chapter 10.

Criteria for Selecting Tests

The previous subsection suggested a general framework within which to assess Head Start effects in the domains of perceptual-motor, cognitive, and language development. This subsection offers some additional preference rules that should govern choice of measurement instruments. These rules were applied in the selection of the battery of measures we propose for the planned evaluation.

Rule 1: Tests must have content validity in measuring a significant area of child performance responsive to Head Start intervention; they also must have content validity in regard to the stated objectives of various Head Start programs.

The instruments finally selected should actually measure what they purport to measure, should measure behavior along this dimension as fully as possible (or at least tap a representative sample of behavior), and should measure areas where it is reasonable to expect program effects. Because the behavioral dimensions receiving preference are those most closely related to social competence, the tests finally selected should have several characteristics that another battery for another purpose would not.

The tests by and large must measure skills relating directly to school readiness or to successful manipulation by the child of his immediate physical and social surroundings. They must be face-valid and amenable to interpretation as criterion-referenced tests, measuring--whenever possible--the same behavior the child will be called upon to perform in the classroom or in some other setting. Although no test is completely atheoretical and every test must in part be validated by an empirical exploration of its concurrent and predictive validity, the instruments in the proposed battery should be as close to being atheoretical and self-explanatory in their significance as possible. A gain or shift in capability on the measures should be readily interpretable as a practical and positive effect in the short term, regardless of its long-term significance.

Measures should also be selected with a reasonable expectation of variation in performance on them among Head Start children, and between the Head Start and non-Head Start populations.

Tests to be administered to the entire national sample should be valid for all cultural groups, although not necessarily with precisely the same significance for each group. The social system and role-specific framework used throughout this report is offered in an attempt to avoid the problems created when investigators assume one or two largely artificial, extreme points of view about the nature of social competence: (1) that it can be generally and publicly defined by psychologists and sociologists, with a clear set of objectively measurable skills representing an equally valid lowest common denominator of capabilities for children of any ethnic or SES group; or (2) that each ethnic, SES, or other culturally coherent social group has its own norms, rules of conduct, and cherished capabilities that make it impossible to speak even of any minimal set of social competence skills appropriate for all children. It is the unhappy lot of the evaluator to try to choose measures allowing comparisons on valid commonalities while allowing various cultural groups a chance of exhibiting a particular dimension of competence as it expresses itself in a familiar cultural milieu. The problem of selecting a few sensible measures with content validity in this sense--accommodating concerns of cultural relativism without requiring different measures for different groups for every outcome--is not trivial. But it is also not impossible.

The selection of measures in the areas of perceptual-motor, cognitive, and language development was conducted with this issue in mind. In general, tests included in the basic battery, to be given to all children in the national sample, should be especially responsive to generalized competencies, although with the prospect of culturally relative interpretation of effects where this seems appropriate. Explicit investigation of culturally unique effects, requiring different measures for different groups, should be the focus of separate studies sponsored by research groups.

Rule 2: Tests must be able to be administered with high reliability in the field.

If an instrument, however attractive, has no reasonable expectation of being administered reliably in the field, it should be eliminated as a candidate for the Head Start battery. In the past, national evaluations of the program often have suffered because instruments were selected for general use *before* it was ascertained that they were not reliable in the Head Start testing situation. Many measures have demonstrated such low levels of reliability when data tapes are examined that they simply have not been analyzed, resulting in enormous waste. Data collection efforts have been pointless because at the next phase of the evaluation the tests have been found uninterpretable.

All aspects of reliability must be high in any new evaluation; test-retest, inter-rater, and--when appropriate--alternative form reliability. This is especially important because many tests will be given under less than ideal circumstances and by paraprofessionals. There are obvious advantages of community involvement and also of cost if paraprofessionals can be used to administer the battery, but such benefits reduce to nothing if in the process the trustworthiness of results is traded away.

One way to enhance reliability is to *select tests that are easy to administer and score*. This test aspect almost has the status of a rule in itself: All measures should be able to be given in the field without complications, they should not take too long, and children's responses should be recordable by simple codes. This is not simply an administrative convenience but is necessary to obtain reliability and validity. Reliability and validity probably increase in such testing situations as a direct function of ease of administration.

Rule 3: The interpretation of children's performance on the instruments selected must be policy-oriented.

The uses of evaluation findings are complex. In many instances the significance of findings is in the eye of the beholder. Choice of certain tests rather than others cannot fully solve this problem, but at a minimum the test selection process should be conducted in a manner mindful of likely policy-related interpretations of outcomes. First, where possible, test results should be interpreted so that they can be of value and interest to multiple audiences--federal officials, Head Start staff members, parents, and elementary school teachers. Although in its principal emphasis the proposed evaluation is summative, not formative, results should inform Head Start practitioners about what they are doing well and poorly, just as it should inform officials in Washington about global program effects.

Making overarching conclusions about Head Start benefits is difficult. Many competing conceptions of success and failure exist side by side, defying easy decision rules about how large a gain must be before it is taken as an indication of children's progress, about how many measures a child must show gains on before one may conclude that he is being affected by the program, about how many children within a program must show improvement before the program itself is declared effective, or about how many centers must be rated successful before the program as a whole is regarded positively. But one thing is certain: *unless instruments have the potential to contribute to policy decisions, they should not be included in the battery.*

This criterion is an easier one to fulfill than the previous two, because no measure accurately tapping some aspect of social competence is without policy significance. But in making marginal choices between measures it sometimes happens that one appears to yield more practical, policy-related information than another, usually because it yields information relevant to differences between program types or it measures an aspect of program effect important enough to influence federal-level decisions about which programs to fund and which not to. Where instruments are more likely than others to yield such information, they should be given preference.

This criterion is particularly clear in reference to measures of minimal information and skills that every Head Start program should successfully impart--"floor" considerations. If it is apparent that an instrument measures minimal knowledge or skills, and that failure to attain these skills is a strong reflection of program inadequacy or strongly predicts school failure, the instrument is especially valuable. Measures of *optimal* Head Start attainment, valuable for different reasons, are not of the same order of policy significance. In this sense, *criterion-referenced measures, or those amenable to interpretation as criterion-referenced instruments, are preferred* wherever the criteria in question are minimal for competence in subsequent schooling or day-to-day social interaction.

Finally, to be of policy relevance, *few tests should be administered.* This may seem incorrect if it is reasoned that it is always useful to have as much information as possible, being selective only later at a stage where there is greater certainty about what is valuable and what is not. But we have learned from past Head Start evaluations that the main problem facing those analyzing, interpreting, and basing judgments on the data is one of too much information, not too little. There has been a tendency to include as many tests as possible in the battery, in part because various factions have sought inclusion of pet measures and accommodation of their concerns seemed the easiest short-run solution, in part, too, because there seemed to be a wise conservatism in hedging bets about where Head Start would have its largest effects, with the chances of effects appearing on some measure more likely when more measures were included in the battery.

This approach has in general been unsuccessful. It has resulted only in much unanalyzed data, lower quality for the data that have been collected (fewer resources have been concentrated on the valid and reliable administration of those measures chosen), and lack of clarity about which measures should be taken seriously at a later stage of analysis and interpretation. Policymakers finally have been presented with an analysis based on three or four child performance measures selected from the much larger set of measures. For all practical purposes it would have made more sense to administer only a few measures from the start.

The lesson to be learned is that it is wiser to make difficult choices early rather than late and wherever possible to test clear hypotheses. We propose a somewhat reduced battery in the realm of perceptual-motor, cognitive, and language development, running the risk that some will feel the battery does not explore all aspects of program-related effects in the three domains. This seems wiser than what we feel is the greater risk of trying to please all factions, offering a compromise potpourri of many measures, and leaving difficult questions of priority among them for later. Policymakers will be grateful for the simplicity and coherence of such an approach, as long as the battery is fair and it samples important areas of behavior in representative domains.

CIRCUS INSTRUMENTS AND THE NATIONAL SAMPLE

The criteria we have used to define important behavioral dimensions and desirable test characteristics considerably reduce the pool of candidate instruments for the new Head Start battery. In the domains of perceptual-motor, cognitive, and language development it was necessary to find tests to measure practical, competence-related outcomes, especially school-related outcomes. Instruments used in past evaluations generally fail to fulfill all of the desired criteria; they lack content validity to assess social competence, sufficient reliability to be interpreted with assurance, or policy relevance.

Fortunately, staff members at the Educational Testing Service (ETS), with experience derived from development of the "Sesame Street" measurement battery, the Longitudinal Study of Disadvantaged Children, the Summer 1966 Head Start "evaluation," and other related projects, recognized this lack of practical preschool outcome measures and several years ago began developing such a set of measures. The resulting battery, "CIRCUS" (ETS, 1973), consists of short, easy-to-administer tests (14 to be administered on a selective basis to children, three that report and rate children's interests and test-taking behavior and the educational environment). The battery is largely completed, although some of the individual tests are currently still under revision; validity and reliability data are available on most instruments in the battery. Complete technical reports are now being made available with data reported for subgroups identified by age, SES, ethnic group, region, sex, and previous preschool experience.

The rationale behind development of CIRCUS coincides well with that of the present evaluation of social competence, and many of the CIRCUS measures are the ones that best fulfill the criteria for instrument selection enumerated in the previous sections. CIRCUS seems to touch upon 22 of the 29 competencies listed by Anderson and Messick (1974); that list is the synthesis of ideas generated at a major conference held for the purpose of defining the dimensions of social competence in young children. The CIRCUS tests are designed as face-valid, criterion-referenced instruments measuring outcomes that preschools have a good chance of influencing and are relevant to the child's competence. In addition, the tests--which can be administered by paraprofessionals--show very adequate levels of reliability in field testing. Marshall Smith (1973, p. 41), the principal analyst of the 1969-70 and 1970-71 cohort data from the Head Start Planned Variation (HSPV) Study, has this to say about CIRCUS:

> A group of us recently completed a series of reports on Head Start Planned Variations, a large-scale field study which examined the effects on children of a number of different preschool curricula. One primary concern during the early planning of the study was to put together a battery of existing tests that would faithfully represent the variety of objectives suggested by different preschool curricula ranging from the Open Classroom type such as Bank Street to academically oriented curricula such as Englemen-Becker. Although we made great efforts to construct an appropriately comprehensive battery, we failed. Almost all of the chosen tests turned out to be close cousins of the standardized achievement test and many were extremely difficult to administer on a large-scale basis. Had the CIRCUS battery existed in a field-tested and reliable form, we might have been able to take giant steps toward the solution of our problems....
>
> The use of a common format across the tests, the focus on ease of administration, and the emphasis on making the tests fun for children would have made the job of administering a battery of tests to 4,000 children faster, cheaper, and far less onerous. These are not trivial points--a single-battery administration in Head Start Planned Variations cost at least $150 and took roughly two hours.

Since the outcomes included in CIRCUS were selected and test items developed according to the experience and advice of teachers and development experts with knowledge of preschools, the battery has a rather different origin and purpose than most other available tests. In the past, Head Start evaluations have been forced to rely heavily on IQ measures and normed achievement tests, many of dubious relevance in the Head Start measurement situation. The behavioral specificity and practicality of CIRCUS is refreshing in this regard, as Boyd McCandless (1973, p. 39) has noted:

> CIRCUS is a test of behavior--how children attack problems--rather than one more extension of testing experts into trait theory. The CIRCUS team members are not looking for an overriding single predictive score, such as the IQ, but rather are sampling behavior in a number of ways so as to guide teachers into diagnostic instructions. My own experience with inner-city teachers of poor black and white children has indicated that there are two major means by which psychologists can help teachers: through their knowledge of (1) principles of behavior management and (2) diagnostic teaching. CIRCUS seems to be a first-rate gambit for giving teachers guidelines for the latter.
>
> CIRCUS is based on the difference, not deficit, hypothesis of children's development and learning. Children are *not* simply lower or higher than one another along a trait dimension of, for example, IQ. They are different. Some solve problems, talk, and think in different ways from others. The different ways are not necessarily better or worse than each other, although they may vary in efficiency. CIRCUS is designed to tap such differences, not to tell a teacher that one child is inferior to another. This is a valuable evaluation concept, or considerably more practical value than testing based on trait/deficit theory.

The CIRCUS tests are by no means the only ones we considered for the new Head Start battery. Many other instruments have been examined, and in some cases those tests have been preferred. But in selecting instruments for administration to the national sample, we make a clear decision to give a prominent position to the new CIRCUS battery. Hence, a variety of the CIRCUS tests form the core of the basic battery suggested in the perceptual-motor, cognitive, and language domains. We do not recommend selecting only one or two CIRCUS tests and then basing the rest of the battery on a piecemeal assortment of other tests or subtests scavenged from other tests. CIRCUS has been designed as a comprehensive and carefully orchestrated set of measures, and joint analysis of a fair number of its various components assures a homogeneous and global assessment of the child. Children appear to enjoy working on the test tasks, and preliminary practice materials and sample items for each of the tests are provided to make sure children understand what they are being asked to do. In addition, the battery's ease of administration is a clear advantage in such a large-scale testing effort.

Pilot testing with the battery indicates that reliability and validity is not reduced in the testing situation by administering CIRCUS instruments to three to five children at once. Rand recommends that this possibility be explored in the national evaluation.

Focused Research Studies

Among the measurement areas mentioned below, a few have been recommended only for focused substudies. Because of high cost of test administration, relevance of results to only a limited subgroup, early stage of test development, or selectivity of content to be measured, it has made more sense to limit testing in these areas to a smaller group of Head Start children.

Some tests simply are too expensive to administer to all children. One example in a domain outside curricular effects is the physical examination to be administered for a special study in the area of children's health and nutrition. It obviously would be desirable to give every child in the national sample a complete physical examination, but in this case the recommendation is that only a small number of Head Start children receive such examinations; the cost of giving full exams seems to outweigh the utility of the enormous amount of information yielded. Analogously, in the area of perceptual-motor, cognitive, and language development it would be interesting to administer a complete battery exploring the child's motor development, collecting large quantities of data, for instance, on fine motor skills. This is simply too expensive and elaborate a venture to propose for the entire Head Start sample, considering the likely payoff in evaluation results.

Although we have argued the case for a common battery for the selected outcomes in the basic evaluation, in certain instances different measures must be used for different groups. Specifically this is true of language measures for Spanish-speaking children where various proficiencies in their own language are the outcomes of interest parallel to those being tested for children whose native language is English.

The third criterion--stage of test development--is particularly constricting. There are several potentially fruitful areas of Head Start testing where measures are still in the experimental phase, or where developmental theory is still too scant to have led to established paradigms and approaches to measurement. One good example of such an area is metacognitive competence--self-awareness of what an individual knows, needs to know, and uses as conscious strategies for obtaining additional knowledge from various sources. The potential payoff of assessment in this area is great, but related measurement technology is still in its infancy. Hence, intensive work best carried out through separate focused studies must be undertaken.

Instruments are also recommended for focused studies if the content they assess is quite specific, of interest only in a circumscribed domain of research or intended to answer limited hypotheses. An example would be measures to test for age shifts among children in information-gathering strategies on a particular classification task, while looking for systematic differences in strategies between cultural groups. Results would be of special interest to Piagetian theorists. Hypotheses on this order of specificity often are as interesting as or more interesting than those testable under a more generalized measurement strategy, but because they are more fine-grained and apt to have circumscribed policy implications, they are not appropriate for exploration in the basic battery.

In making judgments about Head Start success, evaluators are well advised to look at both national and focused studies and to weigh them approximately equally in making judgments about the program. Each represents a rather different *kind* of information about program outcomes; each in its own right is important and useful for deciding "whether Head Start works."

Preferred Instruments

In the discussion below a standard format is used to describe each behavior dimension. First the dimension is described, and then its relevance to the measurement of social competency is considered. Then the candidate measure is presented, with a brief rationale and description of the instrument itself. Finally, there is a brief summary of test characteristics from the ETS CIRCUS pilot study offering basic data about mean scores, reliabilities, and concurrent validity estimates for pilot study samples of kindergarteners and nursery school children.[1]

---

[1] It is important to recognize the strengths and limitations of inferences about test characteristics based on the ETS pilot study. The study was performed with two separate, stratified probability samples, one of kindergarteners (N = 1930) and one of nursery school children (N = 946). School sending areas were randomly selected nationwide from within population density strata, and all kindergartens or preschools within a chosen sending area who consented to participate were sampled, with no more than ten children selected from any one school. The sample gives reasonable baseline data on the CIRCUS measures, but for three reasons its value in predicting test characteristics for the Head Start-eligible population must be qualified.

(1) The two samples, kindergarten and nursery school, are not strictly comparable to each other. In the nursery school sample there are proportionally fewer minority group and low SES children, and proportionally more children with previous preschool experience. In addition, the kindergarten group was tested early in the school year and the nursery school group was tested at mid-year. These differences between samples and testing times help explain the fairly high scores of the younger group in relation to the older group on most measures.

(2) The groups were chosen as a national sample representing all SES levels, with the result that they include small total numbers of low SES and black children. In addition, only the first two tests in the CIRCUS battery were administered to all children; the remainder of the tests were randomly allotted to testing packages such that no child took more than a third of the total battery. For tests other than CIRCUS #1 and #2 only three or four hundred children were in the sample, and further subgroup analysis for low SES and minority children becomes difficult or impossible.

(3) As yet, little data on reliabilities and concurrent validity estimates for low SES children alone are available even for CIRCUS #1 and #2. Many of the statistics presented in the following test descriptions were necessarily based on the entire sample.

With these three caveats in mind it is nonetheless valuable to look closely at what has been learned from the pilot study. The study is far from perfect for our purposes, but even given the problems mentioned above, data on the tests may well prove robust in predicting the performance of Head Start children, especially where cell sizes allow inferences about the low SES subsample.

ETS is compiling complete pilot study data broken down by age, sex, SES level, ethnicity, region, and previous preschool, and information on preferred reporting schemes for the CIRCUS instruments.

Behavioral dimensions are grouped for clarity of presentation in the three areas of perceptual-motor skills, cognitive skills, and language skills. Each of the areas has several subdivisions or subcategories. Preferred instruments are listed within each subdivision or subcategory. The final typology is thus:

Perceptual-Motor Skills:

- Visually guided fine motor skills

Cognitive Skills:

- Global maturational indicator
- Visual recognition and discrimination
- Letter and number recognition
- Problem solving
- Quantitative concepts
- Metacognitive competence (for separate, focused study)

Language Skills:

- Vocabulary development and knowledge
- Comprehension and recall of oral language
- Competence in language use: structured situation
- Competence in language use: unstructured situation (for study with a subsample of the national impact evaluation)
- Metalinguistic competence (for separate, focused study)

PERCEPTUAL-MOTOR SKILLS

In general, many of the dimensions of growth in the area of perceptual-motor skills involve substrata of maturation-related abilities. The issue paper commissioned for the Rand meeting on outcomes and measures in this area and several of the panelists disputed the likelihood of notable Head Start effects.

The panelists felt that there was one reason why perceptual-motor skill development could not be omitted from any evaluation of social competence in Head Start children. Perceptual-motor skills are closely related to the learning of many tasks required in the primary grades; most important, they are prerequisite to learning to write and to the development of proficiency in many concrete operational manipulations.

In addition, many teachers expect competence in certain fine motor skills, such as cutting, pasting, or tying shoes. Children behind their peers in such skills are often perceived as not being ready to learn, or as less than able.

The dimensions of fine motor competency related to school readiness are a priority area of Head Start measurement. Although development of fine motor skills in most children follows a predictable and maturation-related course, the rate of development and mastery are not the same. A lag in skill attainment, especially when a child has arrived at the point where he is physiologically capable of mastering the skill, can become important when the competence in question is a prerequisite to fulfillment of classroom assignments. In addition, many perceptual-motor skills are basic to later operational skills, suggesting that late attainment puts the child at a disadvantage.

Our choice of behavioral dimensions and measures reflects the conviction that assessment should concentrate on criterion-referenced data on the rate at which basically intact children achieved behavior patterns conforming to cultural norms for their age group (i.e., the social system model). Hence, it does not include dimensions or instruments oriented to early assessment of permanent learning disabilities or fine motor anomalies. Another choice we advocate is to pay special attention to visually rather than nonvisually guided fine motor skills. Though the latter are obviously important, too, they tend to be neither as complex nor as directly related to critical reading and math skills.

The preferred measure for visually guided fine motor skills is CIRCUS No. 4: Copy What You See. In an evaluation of social competence striving for economy and representativeness of measurement, *we recommend that major emphasis in the perceptual-motor domain be given to the measurement of hand-eye coordination at school-related tasks, since a major focus is on the child in the role of pupil.* Various instruments, especially IQ measures, include hand-eye coordination items; this has long been considered an important area of measurement not only on tests of school-related achievement but also on basic ability inventories. But hand-eye coordination items on intelligence tests, such as the ones found on the WIPPSI or the Stanford-Binet, are less satisfying as potential Head Start measures than the CIRCUS instrument, Copy What You See. The CIRCUS measure has clear school-related face validity; the task itself, copying numbers and letters, is the same one the child will be asked to perform in school situations.

Jungeblut (1973, p. 26) describes the intent of the test:

> The production of open and closed forms can be discerned from the time of the child's first scribblings, but the preschool child should be able to reproduce or copy from a visually presented form in a controlled manner. In *Copy What You See*, this perceptual-motor coordination is assessed through the child's ability to reproduce such capital and lower case letters as X, P, f, and B and such numerals as 2, 7, and 5.
>
> The child is asked to copy letters and numbers in the bottom half of a box from examples in the top half of the box. Letters and number items alternate. Children are allowed to work at their own pace, and testers offer guidance only in shifting from page to page until the child has completed all 15 items.

The test asks the child to copy stimuli including capital letters (H, B), lower case letters (g, f), forms that are either capital or lower case letters (u, x, k, p, w), numbers (7, 8, 2, 5), and forms that are capital, lower case, and numbers (0, 1). There is no requirement that the child recognize or name letters or numbers. Instead, scoring is based solely on the child's capacity to copy the form precisely; the fifteen letters and numbers are classified according to their component forms: circular (0, 8), straight (single line - 1; two lines - 7, X; three lines - K, H; four lines - W), and combinations of circles and lines (g, p, f, 2, 5, B).

In the ETS pilot study, low SES nursery school and kindergarten children were generally able to complete the 15 item test fully (the mean number of items completed for nursery school children was 15.0, for kindergarten children 14.9). For the full samples, internal consistency and split-half reliability estimates were as follows:

	Alpha	Split Half
Nursery school	0.90	0.90
Kindergarten	0.87	0.89

In addition, to determine the consistency with which scorers followed scoring guides, three scorers rated the same set of booklets twice. Intra-scorer reliabiliities ranged from 0.91 to 0.93; Inter-scorer reliabilities ranged from 0.80 to 0.86.

For the full samples, outcomes on Copy What You See correlated most highly with outcomes on the following other CIRCUS instruments: How Much and How Many (quantitative concepts), Look-Alikes (visual discrimination), Finding Letters and Numbers (letter and number discrimination), Listen to the Story (listening comprehension), and Think It Through (problem-solving):

	Nursery School	Kindergarten
How Much and How Many	0.53	0.53
Look-Alikes	0.43	0.50
Finding Letters and Numbers	0.43	0.47
Listen to the Story	0.60	0.52
Think It Through	0.47	0.49

In the CIRCUS pilot testing there was a 0.17 correlation between sex and total score on Copy What You See, which is the highest for any of the CIRCUS measures. In addition, it was noted that the difference between kindergarten and nursery school children's performance was greater on this instrument than on many others, with kindergarten children receiving substantially higher scores, especially on the more complex configurations. The test did not show floor or ceiling effects for the youngest or oldest children in the pilot study despite the clear age trend. Age-related variability in test performance for the Head Start sample helps provide a developmental baseline against which to measure the results of program-related training in perceptual-motor skills. The teaching of visual discrimination, motor coordination, and visual motor skills has been linked with enhanced performance on subsequent tests of perception, writing, and reading (Campbell, 1971; Lipton, 1969; Maccoby, 1968; Pascale, 1970).

COGNITIVE SKILLS

In the past, cognitive skills have been weighed heavily in gauging Head Start program effects, although because of the global instruments used in previous Head Start evaluations it is probably fairer to consider that past definitions of "cognitive" correspond more or less to perceptual-motor, cognitive, and language skills in the present typology. There is a point, we believe, in separating the strictly cognitive from the perceptual-motor and the linguistic. By so doing, evaluators are more apt to make fine distinctions and are less apt to gravitate to global or unitary notions of skill acquisition.

In the perceptual-motor and language domains, the advocated course is not to select instruments unless they measure malleable aspects of skill attainment, closely related to what Head Start can teach and what the child can learn that is useful in subsequent schooling. In general, this preference also exists in the cognitive domain, although, as will be seen by the first measure proposed, we recommend one departure from this preference scheme by selecting one measure as a maturational marker. The following list of measures include the single maturational indicator and then several measures of constituent dimensions of cognitive skill attainment. All but the final instrument listed are recommended for inclusion in the basic battery.

A Maturational Indicator

Preferred Instrument: Adaptation of Ravens Colored Progressive Matrices

The need for one instrument that will give stable baseline data about the child's cognitive level is particularly important at pretest, when children's entering capabilities need to be evaluated as a backdrop against which to measure improvement during the Head Start year. In the past, no test has been administered in Head Start evaluations expressly for this purpose. However, several IQ measures, notably the Stanford-Binet, have been used as generalized pretest-posttest gain measures, as though they were suitable as outcome measures. *We recommend strongly that the use of IQ tests as pretest to posttest gain measures not be continued.* The measures were not devised for this purpose, they were not intended to be susceptible to program effects, and interpretation of gains is of doubtful validity. Global IQ instruments also tend to include such a diversity of tasks that they have too few items in each component performance realm. In addition, such tests are frequently open to charges of ethnic bias.

The use of baseline data as a *moderator* variable is quite different. Perhaps administered only once at pretest, or perhaps twice as a way of estimating nonprogram-related shifts in the interval of the Head Start year, a single cognitive ability measure to assess basic process is valuable. We recommend that one and one only be administered, and that it serve in subsequent analysis only as an independent variable, a covariate, or other control. It should be treated the same as demographic, SES, and age data.

Among various candidate measures in this domain, the least culturally biased and heavily language-dependent test that connects with a substantial research literature and whose psychometric properties are well known is the Ravens Colored Progressive Matrices. The test is easy to administer and should serve well. Originally adapted from an elementary school version for five- and six-year olds, it will require further age-adaptation for the Head Start population. Both SRI and ETS researchers have developed preliminary versions of preschool adaptations of the test, but a standardized preschool form needs to be pretested.

Visual Recognition and Discrimination

Preferred Measure: CIRCUS No. 3: Look-Alikes

One of the skills most often taught in the preschool, also closely related to later learning in the early elementary grades, is the perception of similarities and differences. Again, the task is a familiar one on preschool tests of ability and achievement: The child is asked to choose from an array of three pictures or figural representations the one that is identical to the model at the top of the page. Many previous tests--the Peabody, the PSI, and the WIPPSI--have items of this sort. The "Sesame Street" test battery also includes such a task.

Look-Alikes, the CIRCUS instrument to test visual recognition and discrimination, includes numbers and letters as well as pictures, enhancing its face validity as a predictor of skill attainment in early elementary school. Jungeblut (1973, p. 25) describes its function as follows:

> It is admittedly difficult to differentiate between perception and cognition. However, traditionally, perception has been defined as the cognition of form, and we are therefore concerned with assessing the visual discrimination and recognition skills that are ordinarily basic to later competency in reading. The *Look-a-likes* instrument samples the child's ability to match to a standard. Both open and closed figures are appropriate at the preschool level, and it is important that the child perceive a unit or form as separate from its background and discriminate among similar units and forms even under simple transformations. For example, in matching to a standard, the preschool child should be able to discriminate among such numerals as 6, 9, and 8 using 6 as the stimulus and among such lower case letters as b, n, and h using an h as the stimulus. In *Look-a-likes*, the child's ability to match series or groups of forms, objects, letters, and numerals is also assessed.

The tester indicates, in turn, the single upper picture or figure, the three lower pictures, and the upper picture again. Then, using natural gestures, the tester says, "Look at this picture in your book... and these pictures. Listen carefully. Mark the one here that looks just like this one." When children begin to understand what is required it may only be necessary to say, "Now we'll do this one," or "turn the page again." In all, there are seven geometric shape arrays, nine letter arrays, six number arrays, and four arrays of familiar objects. This test does not introduce apparent cultural bias in the figures it presents. Among the arrays of familiar objects, all could be expected to be roughly equal in familiarity to children regardless of cultural group or SES level. The test also involves only minimal levels of linguistic comprehension as a prerequisite for task-performance.

The four types of stimuli are varied in their presentation according to orientation, distance, and position. In addition to the total score indicating number of correct responses, a reversal score can be obtained to tell how many of the incorrect options the child selected were reversals of the stimulus presented. Eleven of the stimuli contain a reversal option.

The results of the ETS pilot study indicate that for the pooled national sample of children average performance levels, especially for older children, are very high. The average score correct for both the four- and five-year-old samples was over 75 percent correct, and both groups had almost no difficulty attempting all items on the test. For the Head Start-eligible population we would suspect somewhat lower baseline levels. Low SES nursery school children averaged 16.1 out of 26 correct, and the mean number correct for low SES kindergarteners was 18.8.

Reliability estimates for the total samples are as follows:

	Alpha	Split Half
Nursery school	0.84	0.88
Kindergarten	0.84	0.86

Outcomes on Look-Alikes tended to correlate most highly with outcomes on What Words Mean (receptive vocabulary), How Much and How Many (quantitative concepts), Finding Letters and Numbers (letter and number discrimination), Listen to the Story (listening comprehension), and Think It Through (problem-solving):

	Nursery School	Kindergarten
What Words Mean	0.45	0.46
How Much and How Many	0.69	0.60
Finding Letters and Numbers	0.50	0.46
Listen to the Story	0.65	0.63
Think It Through	0.69	0.59

In the sample of nursery school and kindergarten children, correlations with age were moderate and positive (0.22 in nursery school and 0.13 in kindergarten) and correlations with sex were negligible. Head Start evaluation designers should note that previous preschool experience was more highly correlated with performance on Look-Alikes than on most other measures in the CIRCUS battery (similar or somewhat higher correlations were obtained in both the kindergarten and nursery school samples on Think It Through and Listen to the Story).

Letter and Number Recognition

Preferred Instrument: CIRCUS No. 5: Finding Letters and Numbers

The school readiness literature is full of studies relating children's knowledge and recognition of numbers and letters to their preparation for formal schooling and their later achievement in school, especially in reading. Whether or not the teaching of letters and numbers in the preschool enhances a child's later capacity to read, it has strong face validity as a preschool goal in the eyes of most teachers and parents.

Finding Letters and Numbers is a straightforward measure asking the child to select from among three choices the letter or number named by the teacher. The 15 letters (nine upper case and six lower case) presented in the test were chosen for their frequency of occurrence in English and their configurations--straight lines, open and closed curves, and combinations of straight lines and curves. Five numerals also are included. Items are increasingly difficult, with differences between choices on earlier items in the test requiring fairly gross discriminations (e.g., between J, I, and T), and on later items relatively fine ones (e.g., between p, b, and h). The test may be analyzed by total score or by subscale (upper case, lower case, numerals).

In the ETS pilot study, Finding Letters and Numbers was not difficult to complete for the four- and five-year-olds sampled; mean number of items omitted or unscorable was 0.45 for nursery school children and 0.47 for kindergarten children. Average number of items correct for the total samples were 15.48 out of 20 for the nursery school children and 14.16 for the kindergarteners. Small sample size for low SES nursery school children makes estimates for this group impossible, but for the 133 low SES kindergarteners, mean number of items correct was 12.20. This suggests that ceiling effects are unlikely for the Head Start-eligible population, and that if Head Start children tend to score extremely well on the measure it enables the instrument to be interpreted as a criterion-referenced measure of program-related effects. The instrument also enables comparisons in a very specific range of competency with the outcomes of Sesame Street, which has as one of its explicit goals the teaching of letters and numbers. It would be interesting to analyze Head Start effects on this instrument controlling for degree of Sesame Street watching in the Head Start center, to begin to understand how much new found competency in letter and number recognition can be attributed to the television program and how much to Head Start classroom teaching.

Reliabilities on Finding Letters and Numbers for the total samples were as follows:

	Alpha	Split Half
Nursery school	0.86	0.86
Kindergarten	0.86	0.87

Concurrent validity estimates comparing performance on this test and others in the CIRCUS battery suggest that correlations are highest with What Words Mean (receptive vocabulary), How Much and How Many (quantitative concepts), Look-Alikes (visual discrimination), Copy What You See (perceptual-motor coordination), Listen to the Story (listening comprehension), and Think It Through (problem-solving):

	Nursery School	Kindergarten
What Words Mean	0.36	0.49
How Much and How Many	0.55	0.62
Look-Alikes	0.50	0.47
Copy What You See	0.43	0.47
Listen to the Story	0.18	0.49
Think It Through	0.31	0.52

Strongly differing coefficients for the nursery school and kindergarten samples may to some extent be a reflection of higher reliability in test administration for older children or more tendency for the tests to measure a single latent factor with this group.

Problem-Solving Skills

Preferred Instrument: CIRCUS No. 13: Think It Through

Problem-solving skills are of clear and face-valid importance as they relate to school performance; they also have strong theoretical importance as they relate to various emergent cognitive processes. Many believe that problem-solving skills are the most valuable thing the schools can teach.

The CIRCUS No. 13 test, Think It Through, has the virtue of simultaneously tapping a number of constituent areas of problem-solving (Ekstrom, 1973, p. 27). The test

> is designed to assess five essential abilities: (1) the ability to detect the problem, (2) the ability to define the problem, (3) the ability to use order and sequence in problem-solving, (4) the ability to evaluate possible solutions, and (5) the ability to use classification skills in problem-solving.

The test involves observation acuity, rudimentary notions of causality and inference, classification skills, identifying the first event in a sequence, and evaluating problem solutions. In the first section of the test (six items) the child is shown three pictures and is told, "One of these pictures has something wrong. Mark the picture that has something wrong." An example picture, for instance, shows a wagon which is correctly drawn except for one square wheel. The child must choose the picture with an anomalous characteristic.

For the next items on CIRCUS No. 13, the child is told, "one of these is different. It does not go with the others. Mark the one that does not go with the others." As Ekstrom (1973, p. 28) states,

> During the development of this test, we decided that measuring a child's ability to define a problem could best be accomplished by means of a relational-implicational reasoning type of task that would first require the child to develop the concept of a class from an array of objects and then ask him to select an object that does *not* belong to that class. Rational-implicational reasoning is one of three kinds of reasoning which nursery school and kindergarten objectives frequently mention.

The sequencing notion is again pursued in the next four items, where the child is told, "these pictures tell a story. Mark the one that happens first."

One strength of these sequencing items is that they do not depend heavily on short-term memory (Ekstrom, 1973, p. 29):

> Tasks asking young children to remember sequences have appeared in a variety of tests for this age group. However, this section of the CIRCUS problem-solving test is different because it it not primarily dependent upon short-term memory for sequence or order. Unlike the bead-stringing or block-tapping tests that appear in other test

> instruments, the *Think It Through* sequence items are primarily concerned with real-world events, such as drinking a bottle of pop, building a house, or going down a slide. The child who has observed or taken part in such activities can, of course, solve them by resorting to memory (as in the case of the incongruities items discussed earlier), but even without such knowledge he can reach the correct solution through logical analysis.

The remainder of the test requires more complex judgments by the child. The next six questions request that the child figure out what the character in the picture should do to solve a particular problem, selecting one of three solutions as the best one (e.g., "Here's Clarence's shoe. He's broken his shoelace. Which of these shows what he can use in place of the shoelace?"). The final 14 items require that the child look at three objects drawn at the top of the page and on the basis of a single common characteristic match them with the most appropriate of three objects at the bottom of the page. The child is told: "Look at these carefully (top). They go with one of these (bottom). Mark the one *here* they go best with."

Total score on the instrument can be broken into subscores for identification of a problem (items 1-6), sorting and classifying objects by their properties (items 7-9, 19-32), and evaluating and sequencing (items 10-18). In the ETS pilot study, of the total 32 items nursery school children finished an average of 30.0 and had a mean score of 21.5, while kindergarten children finished an average of 30.7 and had a mean of 22.2. For the low SES subsamples, means were 18.7 correct for nursery school children (N = 57) and 20.2 for kindergarteners (N = 169). For the total samples, all three of the subscores correlated highly with total score:

	Nursery School	Kindergarten
Problem identification	0.60	0.59
Sorting and classifying	0.93	0.91
Evaluating and sequencing	0.80	0.79

Reliabilities on the test were high given its length and complexity:

	Alpha	Split Half
Nursery school	0.82	0.82
Kindergarten	0.81	0.81

Think It Through correlates rather highly with most of the other CIRCUS tests, which is understandable since problem-solving skills are also important for performance on other instruments:

	Nursery School	Kindergarten
What Words Mean	0.43	0.63
How Much and How Many	0.66	0.63
Look-Alikes	0.69	0.59
Copy What You See	0.47	0.51
Listen to the Story	0.67	0.67
Finding Letters and Numbers	0.31	0.52

For both samples, multiple regression analysis introducing background variables as predictors of composite score suggested that previous preschool was a fairly strong predictor of performance on the test.

Quantitative Concepts

Preferred Measure: CIRCUS No. 2: How Much and How Many

Another set of essential skills for the child, also of concern to Piagetian theorists, are those related to number concepts and numerical readiness. Upon entering school the child must have a grasp of basic aspects of enumeration, counting, one-to-one correspondence, ordination, comparison, quantitative language, and other rudimentary aspects of quantification. This area has face validity as a school readiness indicator and also is of considerable theoretical interest; many of the prerequisites for operational thinking and for more powerful cognitive strategies underlying all of the child's competencies are related to attainments in this area.

The CIRCUS test No. 2, How Much and How Many, measures various components of numerical competence. It encompasses the best aspects of the ETS Enumeration Task, used in the last year of the Head Start Planned Variation Study and the ETS Longitudinal Study, and goes considerably beyond the earlier measure in the skills assessed. All parts of the instrument involve identification of appropriate pictures, minimizing the effects of mediating language proficiency. The instrument is designed to evaluate the child's global quantitative understanding (Jungeblut, 1973, p. 23):

> Since our focus was on the age range from about 4-1/2 to 5-1/2, in Piagetian terms we are dealing, except in rare instances, with the non-conserving or preoperational child. In attempting to measure quantitative understandings, we were limited to developing group techniques to assess relatively global notions. For example, in the CIRCUS measure *How Much and How Many* the child is asked to mark among three pictures of elephants the "elephant that is largest," from among three pictures of ponies the "pony that is smallest," the "fewest seals," the acrobat with the "short pole," the "clown with the long nose," and to demonstrate his understanding of *most* in the sense of numerosity (which clown has the most balloons) and quantity (which cone has the most ice cream). It is these global notions, according to Piaget, that are the precursors of numerical comparison.

In each item, the child is shown three pictures and then asked to choose the right one on the basis of some quantitative concept identifying it. Criterion concepts are relational (biggest, smallest, most, least, short, long); numerical (the one with *five* horses), inclusive (all, some, none), and depicting one-to-one correspondence (mark the picture that shows just one ice cream cone for each clown). Several items ask the child to state the number just after or before the numbers in a sequence (what number comes next when you count 1-2-3-4-5? Mark the number that comes after five).

Other items test the child's ability to identify same or different numbers in two sets of objects, test the notion "half," and explore the ordinal concepts of first, middle, and last. "More" and "fewer" comparisons between sets of objects are included, and several items asking the child to match a written number with a set of objects corresponding in quantity to it. Finally, on several items the child is asked to compare three sets of objects with a model set and choose the one with the same number of objects.

In general, all items can be loosely grouped under three categories: understanding of "how many things" correspond to a given number or numeral (counting), comprehension of vocabulary used to express relational terms that are basically quantitative (relational terms), and understanding of one-to-one correspondence (numerical concepts).

In the ETS pilot study, children in both the kindergarten and nursery school samples were in almost all cases able to finish the test. Mean number of items omitted or unscorable was 0.95 for kindergarten and 1.87 for nursery school. Average number of items correct was 28.1 out of 40 for the total nursery school sample and 30.5 for the kindergarten sample. For the subgroup ot low SES children, the means were 24.5 and 27.7.

Of the three component subscales (counting, 12 items; relational terms, 14 items; numerical concepts, 14 items) the relational terms items were the easiest, counting items next easiest, and numerical concepts items most difficult. This is predictable, since the last group of items measures more difficult preoperational skills.

Reliability estimates for the instrument are:

	Alpha	Split Half
Nursery school	0.87	0.88
Kindergarten	0.86	0.87

The quantitative test correlates quite highly with most other CIRCUS instruments, probably measuring in part an ability or generalized competency factor common to all test performance. Along with What Words Mean (receptive vocabulary) it was administered to all pilot sample children. These two tests (CIRCUS #1 and #2) are regarded as the core of the CIRCUS battery.

	Nursery School	Kindergarten
What Words Mean	0.57	0.68
Look-Alikes	0.69	0.59
Copy What You See	0.52	0.53
Finding Letters and Numbers	0.55	0.61
Listen to the Story	0.63	0.70
Think It Through	0.66	0.68

Metacognitive Competence

(For Separate, Focused Study)

In the Rand panel on cognition it was felt that one of the most important sets of skills for subsequent achievement, also one of the most difficult to measure given the current state of the art in testing and measurement, was metacognition--the child's awareness and manipulation of his own cognitive skills. This competence includes an awareness of what one knows, needs to know, and how to get needed knowledge from external sources (parents, peers, teacher, books, displays, etc.) and internal sources (memory search strategies). It also includes utilization of knowledge resources through questioning, perceptual search, and materials search. Clearly the Head Start child, who is preoperational, has not attained a large proportion of his ultimate capacity in this realm. Much of the capacity will have to wait for the attainment of concrete or even formal operational thinking. But the preoperational child may show precursors of this skill in choice of information-gathering strategies, choice of strategies to use in solving problems, and so forth.

It may prove, as panelist John Flavell (1974, p. 3) has suggested, that "meta-anything is inordinately difficult for a child of this age to do, even after considerable training." However, certain metacognitive strategies are so clearly adaptive in school situations that any child who has even partially mastered them is at a clear advantage. For instance, children are more likely to be successful in the early school grades if they can monitor personal states of learning and knowledge: knowing how well they understand what is being taught, and making special effort where it is apparent they do not understand. Similarly, it is valuable for children to know who and what constitute the most valuable sources and resources for needed knowledge or skills and how to best use these sources. Another aspect of metacognition has been emphasized by Jerome Bruner in his research: the aspect of problem-seeking strategies, or knowing how to ask good questions and pose good problems. There are also useful skills of selective attention and memory search.

In this important area there are at present few adequate measures, and fewer still that are appropriate to the Head Start testing situation. Kruetzer et al. (1974) offer a basic framework for research on metacognition, and Kagan's Matching Familiar Figures test and Anderson and Messick (1974) offer partial approaches to studying the phenomena. But there is a need for practical measures adapted to the preschool classroom. *We recommend that metacognitive development and learning in Head Start be the subject of a separate inquiry ancillary to the national evaluation.* The topic is deserving of careful research, but as yet hypotheses and instruments that would enable it to be tested are tentative or experimental. An exploratory approach should be permitted within the framework of the national evaluation.

LANGUAGE SKILLS

Aspects of linguistic competence that should be tapped in the Head Start evaluation include (1) the referential function of language, (2) the area of social negotiation skills, and (3) the general use and awareness of language for the self.

The *referential function* of language includes vocabulary, verbal comprehension (the ability to understand increasingly complex language), productive language, and the cognitive reasoning skills closely related to language skills in their attainment. *Social negotiation skills* refer to the use of language skills in real life, in dealings with other children and with teachers. Code switching, the ability to adapt one's patterns of language to the specific context, and the use of appropriate forms of address and appropriate questions are included under this category; practical social negotiation skills enable children to be more effective speakers or listeners in relations with their peers and adults. The third area--*awareness of language and its use for the self*--includes the ego-control functions of language, linguistic and dramatic play, the use of language for self-reference, and metalinguistic understanding (awareness of one's own linguistic strategies for gathering, encoding, and retrieving information).

Rand panelists considering language development agreed that all three of these aspects of language acquisition and competency were important, but they also recognized that, given the state of the art in language skill measurement, the first was easiest to measure directly, the second was next most difficult, and the third lacked techniques developed beyond the experimental stage. We have therefore had to accept certain practical constraints in what it is possible to assess.

Another useful way to think about language development in the Head Start program is by making a three-part distinction among linguistic competence, competence in language use, and the language children actually use--that is, between what the child *can understand*, what he can *produce when called upon* and what he *actually produces in a natural situation*.

In the realm of linguistic competence, Featherstone (1973, p. 17)[1] has commented, "It seems probable that Head Start will influence children's actual performance in particular situations more than it will affect their fundamental linguistic competence." This implies that looking for some change in syntactic knowledge is inappropriate, since--even if one could avoid measurement problems created by dialectical differences--it is not likely that there are program-related differences in language development or developmental differences from group to group. Instead, it makes more sense to explore such areas as increase in vocabulary and increased sophistication in the child's *style* of language use.

Rand panelists recommended first that vocabulary knowledge and increase in vocabulary use be assessed. Just as in other domains, this kind of learning should be measured with instruments of high face validity for preschool and early elementary school situations; items also should be chosen in part because they are likely to differentiate between the Head Start and non-Head Start populations. Measurement should not involve tests with likely cultural bias, such as the Peabody Picture Vocabulary Test. There may be a need to design a few structured situations that depart from the usual adult-child test situation, that are more life-like and interesting for the child, and that tap aspects of language use of practical importance. For such measurement, one procedure is the analysis of referential language in some version of the two-person communication game. More will be said about this approach below.

---

[1] Issue paper prepared for the Rand Language Development Panel: "Assessing Language Development among Head Start Children."

Is the goal of Head Start to impose middle-class American culture and standard English on the disadvantaged? We believe the answer is yes only insofar as children should be enabled to deal with institutions impinging upon them that are manifestations of that culture. To the extent that language is seen as a tool for cognition, non-standard English or the child's primary language (e.g., Spanish) should be used in Head Start instruction and evaluation; but to the extent that the aim is to promote communication between the child and his public school teachers, the program should be assessed as to its effects on the use of standard English by children. Language is thus embedded in a social context that must be taken into consideration during the evaluation, with recognition of various speech communities. One of the most important skills children acquire during the preschool years is the ability to shift and adapt language to the specific social roles they are asked to play, and to the even more specific content demands of given situations. This aspect of sociolinguistic competence is a valuable skill for the Head Start child to acquire, fundamentally strengthening his repertoire of strategies for coping with the larger world of adults and peers. It is also an aspect of a person's cognitive *style*, something that must be learned before the child is effective in obtaining information and fulfilling his goals in the school and neighborhood.

Assessment techniques for the quality of the child's language production in relation to social settings are not nearly as well developed as measures for semantic and syntactic proficiency in traditional testing situations. The social uses of language deserve special attention, but full measurement of actual language use would require systematic observation in the classroom and home, which seems well beyond the capacity of the present evaluation to deliver. Some dimensions of actual performance patterns should be tapped by the measures proposed for the basic battery--those amenable to assessment in one-to-one testing or small-group classroom situations. Dimensions that can be assessed only by observing in structured situations should be the focus of substudies.

Language acquisition studies in this country have focused for the most part on English language acquisition. This is sensible in Head Start evaluation as long as the principal language of the population served is English, or the principal influence and source of competence in the child's life is acquisition of spoken English. If Head Start is to serve the needs of the Spanish-speaking population, however, it should also develop Spanish language acquisition and learning. Therefore, optimally an evaluation would be concerned with the influence of English on children's Spanish as they develop, the social contexts in which children fluent in both languages use each language, and the contexts in which they switch codes. It should also explore the teaching strategies most effective in producing bilingual fluency and the most appropriate teaching techniques for children whose first language is Spanish.

Most of these questions, while important and entirely relevant to Head Start, are beyond the ability of present assessment techniques to answer fully. Unless valid adaptations of the recommended tests are made during the preparatory year, the evaluation must limit itself to a focus on the bicultural child's ability to operate successfully in a standard English milieu, rather than explore as well his ability to gain competence in a native language medium. We strongly recommend, however, that at minimum the three CIRCUS measures suggested below be adapted or used as guides in parallel test development for Spanish-speaking children so that their language competencies can be adequately assessed. We also recommend that no approach to data analysis in the forthcoming evaluation overlooks the valid and equally important relations of various dialects and native languages to children's learning in standard English.

In general, for the basic battery we recommend that two measures be used in the area of semantics--one of vocabulary development and knowledge and the other of comprehension and recall of oral language--and two measures of competence in language use, administered in a structured situation. Two instruments are recommended for competence in productive language exhibited in an unstructured situation, to be used with a subsample. In addition, it is suggested that metalinguistic skills be explored through separate, focused research studies. Descriptions of language domains and instruments follow.

Vocabulary Development and Knowledge

Preferred Instrument: CIRCUS No. 1: What Words Mean

Evidence from past studies indicates that children learn specific vocabulary in preschool. Head Start is expected to make a difference in this area. The best format for measurement assesses the child's understanding of words without requiring that he or she verbally produce synonyms or definitions. The child should be able to point at one of several pictures that are candidate referents and choose the correct one to correspond to the word given. Ideally, there would also be a number of open-ended items in the same format such that the child could choose more than one correct picture corresponding to the word he hears.

The CIRCUS No. 1, What Words Mean, is a good vocabulary measure for Head Start administration. It adheres to the preferred format, and the pictures from which children are asked to select a correct answer are of suitable content validity for various ethnic and geographic groups, a characteristic few other vocabulary tests can claim. The child is told: "Look at these pictures in your book and listen carefully. *Bridge*. Mark the one that is a *bridge*." Vocabulary items are for the most part nouns, but there are also several modifiers and relational terms ("mark the one that shows the *front*"), and several verbs ("mark the one that *floats*"). Some nouns were chosen as representative of classes of words, such as insect, furniture, and jewelry. Secondary meanings or denotations of some verbs also were included, e.g., *pour* (rain), *run* (an open faucet). In all, there are 40 words.

What Words Mean and its quantitative counterpart How Much and How Many are the core of the CIRCUS battery. Because it is a picture vocabulary test, the instrument does not permit definitional elaboration, an aspect of language competency to be discussed further below. But it does include distractors chosen carefully to test the exactness of the child's understanding of the concept or object presented (Tanaka and Nassad, 1973, pp. 15-16):

> [O]ur work with the CIRCUS vocabulary measure represents an attempt to correct a problem that is common in many of the picture vocabulary tests used for this age range. Quite

> often, the items in such tests measure only the child's global understanding of a word. Thus, the distractors have little or no relationship to the target word, and the child need only a vague association with the required word in order to eliminate the wrong answers. In the development of the items in the CIRCUS vocabulary test, there was a deliberate focus on the careful use of distractors that would measure the preciseness of the child's understanding--if the stimulus word was 'log,' the item included drawings of a piece of lumber and a tree as well as a log.

In the ETS Pilot Study for the total nursery school and kindergarten samples the mean number of items omitted or unscorable was 2.60 for nursery school and 1.01 for kindergarten children. The nursery school group averaged 27.8 items correct out of 40, and the kindergarten group 30.1. For the low SES subsamples, the mean numbers of items correct were 26.6 for nursery school and 27.3 for kindergarten. In general, as would be expected, the nouns were easier for all children than the verbs or modifiers.

Test reliabilities for the total samples were satisfactory:

	Alpha	Split Half
Nursery school	0.86	0.87
Kindergarten	0.83	0.84

Correlations with other instruments in the CIRCUS battery were generally quite high, although not as high as for How Much and How Many:

	Nursery School	Kindergarten
How Much and How Many	0.58	0.68
Look-Alikes	0.45	0.46
Copy What You See	0.33	0.44
Finding Letters and Numbers	0.36	0.49
Listen to the Story	0.46	0.70
Think It Through	0.43	0.65

Comprehension and Recall of Oral Language

Preferred Instrument: CIRCUS No. 9: Listen to the Story
Strongly recommended for adaptation to Spanish-American children

Another important aspect of semantics is comprehension, interpretation, and recall of sustained spoken language. The child must be able to assimilate and integrate information as it is heard, and make sense of it. These capabilities can be measured with young children by telling them a story and then asking them questions about it.

CIRCUS No. 9, Listen to the Story, requires the child to look at successive groups of three pictures in the test book. The child is told, "Look at these pictures. I'm going to tell you a story about them." Then, as a first item: "Listen carefully. The children were excited when they saw the sign for the circus. It had a clown on it. Mark the picture that shows the *sign* they saw." The child must choose the correct picture from an array of three. The test includes 25 items, 15 requesting identification of a character or object ("mark the animals they saw") and 10 requiring more complex forms of understanding ("mark what Seeley did first"; "mark the one they cannot have").

Listen to the Story has the additional advantage of being quite easily adapted to non-English speaking or bilingual children. Slight modifications in the existing test could tap semantic competence in the child's native language as well as the second language. A Spanish-American version of the test could be devised fairly easily, resulting in a culture-fair measure of comprehension and recall, something Head Start assessment has never before been able to offer. A truly culture-fair instrument would be of wide interest to Head Start teachers, parents, and officials.

Performance of the nursery school and kindergarten children in the national pilot study indicated that there were minimal numbers of items omitted or unscorable for the two populations. Mean number of items correct for the total nursery school group was 18.0 of 25, and for the kindergarten group 18.9. For the low-SES subsample the corresponding numbers correct were 17.8 and 16.6.

Test reliabilities for the total samples were as follows:

	Alpha	Split Half
Nursery school	0.77	0.78
Kindergarten	0.79	0.80

The relationship between CIRCUS Ns. 9 and the other CIRCUS measures proposed for Head Start testing are quite strong in the national pilot study samples; highest correlations are with What Words Mean (receptive vocabulary), How Much and How Many (quantitative concepts), Look-Alikes (visual discrimination), Copy What You See (perceptual-motor coordination), and Think It Through (problem-solving):

	Nursery School	Kindergarten
What Words Mean	0.46	0.70
How Much and How Many	0.63	0.70
Look-Alikes	0.66	0.63
Copy What You See	0.60	0.52
Think It Through	0.67	0.67

Competence in Language Use: Descriptive and Narrative Use

Preferred Instrument: Sections I and III of CIRCUS No. 10: Say and Tell
Strongly recommended for development: two-person communication game

The realm of language use is fully as important as the realm of language capability. For testing situations, it can be further subdivided into those aspects of use that emerge with structured items and those that can be assessed only with open-ended items. For purposes of the basic battery in the forthcoming evaluation, we believe it would be unmanageable to recommend that measures difficult to code or interpret be given to all children in the national sample. Unstructured items, which we believe are important to use with a subsample, are more costly and more difficult to analyze than structured items or language elicited in a structured setting. We therefore recommend a two-part strategy, with half of CIRCUS No. 10, Say and Tell--the half that can be easily coded and interpreted--to be given to all children in the sample. The other half of the test, to be discussed separately below, should be given only to a subsample.

Tanaka and Massad (1973) offer a helpful description of the components of the test:

> We agree that this real world of language performance cannot possibly be fully explored through the use of any prescribed set of standardized measures. At the same time, there is a need to provide some way of helping the teacher to sample the richness of the child's oral language. *Say and Tell* measures the growth of the child's spoken language by observing three types of language use:
>
> 1. The descriptive use of language: The child is handed a common object and is asked to describe it. One item elicits the child's use of categorical language such as asking for various attributes ("What color is it"?). Another merely asks him to "Tell me all about that."
> 2. The functional use of language: The child is shown a number of pairs of drawings. A statement is made about one of the pictures, and the child is asked to complete the statement that applies to the other picture ("Here is a boat. Here are two _____."). There are 38 items dealing with such things as the use of plurals, verb tenses, prepositions, subject-verb agreement, comparatives, possessives, and so on.
> 3. The narrative use of language: The child is shown a large colored drawing, and the teacher explains that it is a picture out of a storybook, but that "I don't have the story that was in the book, so I want you to make up a story to go with this picture. What do you think the story was about?"

We propose that Parts I and III be included in the basic battery. Part II is recommended only for a substudy, not because it is less important but because it promises to be challenging to interpret (Tanaka and Massad, 1973, p. 18):

> In the measure of functional use of language, many of the responses showed that the children clearly understood the task but were managing it in their own language. For example, in one of the items on verb tenses, the teacher pointed to each of two drawings of monkeys and said, "*This* monkey ate his banana. *This* monkey is still _____." Back came responses such as, "This monkey is still *not finished*." This monkey is still *hungry*." "This monkey is still *chewing*." "This monkey is still *holding his banana*." As a result of this delightful but frustrating experience, we now have a tremendous respect both for the young child's command of his language and for the coding problems of researchers who have been working in this field.

Part I of Say and Tell, recommended for use with the full Head Start sample, is further subdivided into two parts. The first section reveals the child's ability to describe common objects (e.g., a pencil). The child is shown the object and asked, "What is that?" Response categories are: no answer, pencil, pen, other (_____). Then the child is asked six additional questions about color, shape, what it is made of "what you can do with it," "what else does the same thing," and "can you tell me anything else about it?" Answers are scored according to pre-established categories, but these categories have been developed on the basis of the most common variants of the correct answer given by children in pretesting (e.g., pencil and pen are both correct). In the second section of Part I, the child is asked to describe two pennies in as much detail as possible. The description is coded according to the categories of label, class, color, shape, material, function, number or value, other physical characteristics, and comparative characteristics. Probes are permitted to be sure the child has said all he or she can ("Suppose I don't know what pennies are, what can you tell me about them?").

In Part III of Say and Tell, the child is shown a picture in a booklet and asked to make up a story about it, a technique similar to traditional projective testing. Probes again are allowed to be sure the child has said all that he can about the picture. The coding of the stories has two aspects--quantity and quality. Quantity is assessed according to total number of words the child uses, and number of different words; quality is rated according to labeling, verbs, modifiers, syntax, sequence, plot extension, organization, feeling, rhythm and cadence, comparison, character extension, and spatial elements. The more complex the story along these dimensions, the higher the absolute score of the child if all scoring categories are aggregated. Of course, the test also can be analyzed with regard to component categories and with special weighting of certain categories.

Internal consistency estimates on Parts I and III for the two national pilot study samples are as follows:

		Alpha
Part I	Nursery school	0.72
	Kindergarten	0.49*
Part III	Nursery school	0.78
	Kindergarten	0.78

---

*Low reliability for the kindergarten sample on Part I is difficult to understand, especially in light of the more adequate level for nursery school children. This may well indicate a mistake in coding or data analysis.

Inter-rater reliability, of special interest because of the complex coding scheme on CIRCUS No. 10, was satisfactory in the pilot study, never dropping below 0.80. Correlations of performance on Parts I and III of Say and Tell with performance on other CIRCUS measures were not as high as for many of the other instruments, suggesting that the test measures rather a different cluster of skills; highest correlations were as follows:

	Nursery School	Kindergarten
Part I		
What Words Mean	0.36	0.38
How Much and How Many	0.43	0.35
Look-Alikes	0.32	0.15
Listen to the Story	0.51	0.47
Think It Through	0.38	0.37
Part III		
What Words Mean	0.02	0.12
How Much and How Many	0.07	0.12
Copy What You See	0.01	0.17
Listen to the Story	0.19	0.38
Think It Through	0.38	0.24

In addition, the correlation of subscale scores for Parts I and III of Say and Tell also were rather low: 0.31 for nursery school children and 0.24 for kindergarten children. This suggests a considerable difference in what the two parts measure, even allowing for the somewhat low reliabilities recorded for Part I.

One additional measure is recommended for development in the area of language skill assessment to fill out the basic battery.

A member of the Rand language panel[1] summarized panel opinion about the importance of an additional measure in the area of language use:

> It seems extremely important to design a few structure situations which get away from the usual adult-child test situation, which are more intrinsically life-like and interesting for the child, and which tap aspects of language use we think important.
>
> The only such assessment procedure I feel sure can be designed in time for HS use in referential language is some version of the two-person communication game. This can involve giving and receiving descriptions or instructions, in statement form (as in Bob Krauss' versions) or question form (as in Vera John's 2-child puzzle game). The stimuli can be objects and pictures in various domains to tap varied vocabulary and sentence structure. The procedure can be so constructed to assess simple visual discrimination, vocabulary necessary to differentiate the particular stimuli, ability to construct an utterance which combines the necessary attributes, complexity of the mode of sentence construction chosen, and (if feedback from the listener is possible) how the child as speaker or listener deals with the reparation necessary when information is inadequate. Here (as elsewhere) all verbal instructions and stimuli can be in standard English, but answers must be scored so that dialect differences penalize no one.
>
> Of all the suggestions we made for new assessment measures, this is the oldest, has the first research base in terms both of social class differences in language use and of replicated experimental procedures for assessment.

We recommend that the OCD or the primary evaluation contractor immediately discuss the development of such a measure with experts in the fields of language and testing and measurement, and that such a measure be designed and pretested with a low-SES population before the initiation of the proposed evaluation. With the exception of Spanish-American instruments parallel to CIRCUS Nos. 1, 9, and 10, this area is the only one where we believe that basic instrument development is necessary to devise an additional measure for inclusion in the basic battery. This recommendation is strong: We would like to see the measure added because it covers an important area of language competence not otherwise assessed but central to the goals of Head Start.

---

[1] Cazden, C. B. (1973) response to Rand Cognitive Development Panel.

Competence in Language Use: Functional Use and Use in an Unstructured Situation (for subsample study)

Preferred Instrument: CIRCUS No. 10 (II): Say and Tell
Supplemental Instrument: Observational System

Open-ended language testing and assessment in unstructured situations is the other half of measurement in the domain of language use; it deserves study with a subsample of the proposed evaluation. Two kinds of measures are needed: First, it is important to include some individually administered test of functional language use complementary to Parts I and III of CIRCUS No. 10. Second, there is a need for some observational measure of actual language performance and practices in the social world of the Head Start classroom, away from the one-to-one testing situation.

The open-ended Part II of the CIRCUS No. 10 is appropriate for the individually administered component of assessment. Part II of the test provides the child with a sequence of pages with two pictures on each. The child is shown one picture and given a sentence (Here is a *duck*). Then the child is shown the other picture and given a partial sentence to complete (Here are two [*ducks*]). Response categories in this case are: NA, ducks, duckies, duck, other (_____). Answers are coded so that they can be indexed according to semantic correctness, or semantic and syntactic correctness. The words involved are all easy enough so that it is unlikely that the correctness of children's answers will be an artifact of vocabulary knowledge. Words to be filled in are nouns, verbs, and relational terms. Part II of Say and Tell assesses the child's ability to use correct forms of plurals, verb tenses, prepositions, subject-verb agreement, comparatives, and possessives. The scoring scheme assigns two points to an intended response and one point to an unintended but grammatically correct response.

Internal consistency estimates for the present scoring scheme are adequate, but the Office of Child Development may want to propose that scoring for Part II of CIRCUS No. 10 go through successive refinement in the coming year:

	Alpha
Plurals	
Nursery school	0.69
Kindergarten	0.75
Verbs	
Nursery school	0.71
Kindergarten	0.71
Other	
Nursery school	0.70
Kindergarten	0.72
Comparisons	
Nursery school	0.81
Kindergarten	0.81
Total	
Nursery school	0.89
Kindergarten	0.90

Part II of Say and Tell correlates somewhat more highly than other parts of the instrument with other CIRCUS measures. Highest correlations are as follows:

	Nursery School	Kindergarten
What Words Mean	0.31	0.50
How Much and How Many	0.51	0.43
Finding Letters and Numbers	0.41	0.15
Listen to the Story	0.53	0.60

Correlations are higher between Parts I and II of Say and Tell than Parts II and III:

	Nursery School	Kindergarten
Part I	0.52	0.34
Part III	0.18	0.25

In addition to the measurement of functional language, some aspects of language use can be determined only by observation in specified settings. None of the currently available observational measures is fully satisfactory, and we therefore recommend that for the subsample study an instrument be further developed, perhaps along the lines of the observational system proposed by Tizard et al. (1972). The final measure should be pilot tested during a preparatory year prior to initiation of the study; this pilot investigation probably should be the same one in which the "two-person communication game" instrument is pilot tested.

Metalinguistic Competence

For separate, focused study

Just as in the cognitive domain, "meta" phenomena in language development are of special interest but also are especially difficult to measure given the state of the art in developmental theory and instrument development. We therefore recommend that metalinguistic competence be explored in Head Start, but on an experimental basis.

Metalinguistic competence encompasses the ability to analyze language--for example, to correct one's own semantic and syntactic errors, to criticize one's own speech forms, to be aware of the kinds of speech most apt to influence listeners, to adopt conscious strategies for utilization of speech skills, and to use verbal rehearsal as a mnemonic device. Use of rhythms, puns, and other forms of playing with language are also ways in which the child develops his own language competence. Benefits presumably include higher self-esteem for the child, greater adaptability to different settings (code-switching techniques), more access to one's own memory, and more art in presentation of familiar ideas. We do not know whether such competence is likely to be responsive to Head Start treatment, but if it were, it might turn out to be one of the principal benefits enjoyed by Head Start children. Were it not for the absence of measures, this would be an area of major interest for testing in the national sample or a subsample study.

We recommend a series of focused studies of metalinguistic phenomena to be closely coordinated (or an integral part of) the study of metacognitive phenomena. It seems wise to combine the two areas of inquiry; they overlap substantially, and explorations in each area can contribute to understanding in the other.

Assessing Test-Taking Behavior

It is important on all measures proposed for use with the national sample that child performance be assessed for response style and test-taking behavior as well as for correct and incorrect answers. Much of the increased competence of children may be reflected in their approach to answering questions as well as in the increased knowledge or skill demonstrated by answering correctly.

The CIRCUS Behavior Inventory (No. 16) is the preferred measure of response style. (Chapter 5 includes an additional recommendation for this area.) Messick (1973, pp. 36, 37) makes a strong case for its relevance to preschool assessment:

> The context of the measurement process itself is most usefully assessed not so much by documenting objective characteristics of the tasks, the tester, and the situation as by recording the child's stylistic reactions to them. This is usually accomplished by means of direct tester or teacher observations of the child's stylistic responses to the cognitive demands of adaptive requirements of the measurement tasks. These ratings may be made separately for each task, for a representative selection of tasks, or globally for the battery as a whole. *The CIRCUS Inventory of Test-Taking Behaviors* includes judgments of such aspects of the child's responsiveness as:
>
> - the degree to which he asked for help
> - refused or indicated reluctance to work on tasks
> - expressed enjoyment or amusement over particular content
> - indicated that he didn't know certain answers
> - indicated a desire to stop
> - appeared to respond at random
> - appeared to weigh alternatives carefully
> - spoke about or attended to unrelated objects or events
>
> By relating stylistic consistencies in test responsiveness to patterns of test performance, the validity of test interpretation is likely to be improved, regardless of whether these response styles are transient and specific to particular tasks or situations or are more generally characteristic of the test-taking behavior of the subject.

In the proposed evaluation, measurement of test-taking behavior should be an integral part of administering the basic battery. The CIRCUS Behavior Inventory is completed by the teacher or tester after observing the children taking one or more of the CIRCUS measures. In Head Start testing, the Inventory could be recorded after each test, after groups of tests, or once at the end of the testing session. A factor analysis of the Inventory indicates two basic factors, one reflecting children's enjoyment of test content and willingness to follow test procedures and the other reflecting children's attentiveness to tasks and willingness to complete tasks. The reliabilities of the two scales developed on the basis of these two factors for the ETS pilot study sample are high.

	Alpha	
	Scale 1	Scale 2
Nursery school	0.84	0.88
Kindergarten	0.74	0.86

In general, data from the pilot study also established that the majority of children in the pooled sample were able to handle CIRCUS tasks without misunderstanding directions and with minimum difficulty introduced by conditions of the testing situation. The children were reported usually to keep their places on tests, and rarely or never to indicate reluctance to work on tests, ask for help with the test from the teacher or other adult, say answers aloud in group testing, comment on other children's work, answer questions before directions were completed, indicate a desire to stop before the test was over, appear to answer "at random," or become distracted by unrelated objects or events.

Correlations among CIRCUS Behavior Inventory subscales and CIRCUS measures for the pilot study samples were often rather high, confirming the importance of this aspect of measurement:

	Enjoyment		Inattentiveness	
	Scale 1		Scale 2	
	Nursery School	Kindergarten	Nursery School	Kindergarten
What Words Mean	0.31	0.41	-0.34	-0.29
How Much and How Many	0.47	0.47	-0.37	-0.36
Look-Alikes	0.53	0.41	-0.41	-0.31
Copy What You See	0.23	0.36	-0.37	-0.36
Finding Letters and Numbers	0.30	0.35	-0.16	-0.29
Listen to the Story	0.39	0.43	-0.43	-0.36
Say and Tell				
Part I	0.39	0.28	-0.18	-0.14
Part II	0.42	0.25	-0.33	-0.29
Part III	0.21	0.21	-0.04	-0.09
Think It Through	0.54	0.36	-0.19	-0.30

INSTRUMENT SELECTION AND EVALUATION DESIGN

This section will briefly discuss some of the design implications of our recommendations on the perceptual-motor, cognitive, and language development test battery. One important implication is that, for purposes of the national impact study, the same tests will be given to four- and five-year-olds. Although it is clear that some of the tests may have a somewhat different meaning as they measure the performance of younger children and older children, with somewhat different skills and test-taking abilities being tapped at various ages, it is equally clear that for purposes of the national evaluation there is both a theoretical and a practical advantage to administering the same tests to all children. The theoretical advantage is that this approach allows generalizations about the entire Head Start population and continuity of assessment across age levels. The practical advantage is that it is far simpler to administer a single basic battery than to give two or more different batteries for children of different ages.

We have selected tests for the basic battery designed for both four- and five-year-olds. These tests do not have obvious floor or ceiling effects for either age group; the skills assessed also are relevant for Head Start children of various ages. In addition, one essential strategy for interpreting evaluation results is that age be entered in all analysis as a covariate or independent variable, controlling for age-specific effects and resulting, in effect, in comparisons between age-mates. The regression model used by Smith et al. (1973) and Weisberg (1973), introducing age as a continuous variable, is a good example of one legitimate analytic strategy.

A second implication of the choice of the basic battery is that the same tests will be applied across cultural groups. It is the express purpose of the evaluation design (see Chapter 7) to facilitate within-group comparisons for each cultural group, allowing children to be compared with controls from the same cultural background, rather than being concerned with inter-group comparisons. Presumably the measures included in the basic battery will permit within-group comparisons as well for one cultural group as for another. Except for specific cases where it is inappropriate (e.g., Spanish language development for Puerto Rican and Chicano children), all tests in the basic battery will be administered identically and scored identically. In order to minimize tester effects, actual testing should be performed whenever possible by persons of the same cultural group as the children tested. It may also be advisable at the time of data analysis and interpretation to invite members of various cultural groups to propose new ways of looking at the data to yield culturally relevant insights.

Pretests

Two design issues remain that are not resolved in the choice of a basic battery and must be considered separately. The first issue concerns pretests. As already noted, pretests on all instruments probably are necessary for two reasons: first, to establish the differences, if any, between Head Start and control groups in sites where random assignment is not feasible, so that appropriate correctives can be applied in the analytical phase; and second, to be able to conduct analyses on the basis of entering characteristics of children so as to document any possible differential effects. The question is whether to (1) administer the full basic battery at pretest as well as at posttest, (2) use only a subset of the battery, or (3) give all the tests to only a subset of children in the national sample, as suggested for the health and nutrition subbattery. The validity of the latter approaches is, of course, dependent on a decision to randomize children to Head Start and non-Head Start groups and may not appear attractive if for some reason randomization cannot be effected. But assuming that it can, it is fair to ask: What are the relative merits of various partial pretest designs, and what are the losses in not administering the full battery as the pretest? There are three advantages to a partial pretest design: reduction of possible practice effects on children, lowering the possible effect of the battery on the intervention (i.e., Head Start teachers being influenced by what they perceive to be the curriculum expectations of the test battery), and reduced cost.

Practice Effects. Obviously, administering a partial battery would eliminate practice effects for those tests not administered. It might even be possible to vary the specific tests used for the partial battery in different sites so as to give some information on practice effects. Similar information could be derived from an administration of the full battery to a subsample. It is unlikely, however, that practice effects will be much of a problem even if the whole battery is administered as a pretest to the entire sample. First of all, both the Head Start and the control group children will be taking the tests, so that one would have to hypothesize an interaction effect between the treatment and the pretesting for the practice effect to become a concern in the analysis. Second, it is highly unlikely that children will remember much of the test substance in the intervening six or more months between pre- and posttests. There may be, of course, a pure test-taking practice effect aside from substance, but again, this would be the same for Head Start and control groups. Thus, the loss of information incurred by not giving the whole battery to the full sample hardly warrants a partial pretest design just to guard against a probably minimal or nonexistent problem.

Intervention Effects. The possibility that Head Start staff might be affected in their teaching by the content of the test battery is a real one. However, a partial pretest design will not avoid this problem. The reason is that the informational activities that must be carried out before the evaluation gets under way (see Chapter 9) will already provide full knowledge of the test battery. This is particularly true since teachers will be asked to weight sets of outcomes according to the importance they assign to each. Hence, the confounding of the intervention by knowledge on the part of the teachers about the test battery is not likely to be exacerbated by pretesting, so that little difficulty is created by administering the full battery to all children included in the sample.

Costs. Reduction in monetary costs can be effected, at the cost of loss of information, by either partial pretest design. Administration of a partial battery could cut testing days by as much as 50 percent. If we assume 20 person-testing-days per site (encompassing 30 Head Start children and 30 controls) for the full battery--remembering that most of the tests can be administered to small groups rather than individually--at a cost of $25 per day, then costs could be reduced from

$500 to $250 per site by using a partial battery. For 100 centers, this would mean a cost saving of $25,000 for the pretesting. Some tester training costs could be saved as well, though not in proportionate amounts because pretest training will presumably decrease the time needed for posttest training. Even greater savings could be effected by selecting a random subsample of, say, 25 centers for pretesting with the full battery. Clearly, this option is available only if at the other 75 centers a random assignment strategy is instituted. Further, it will severely limit the power of any analysis of differential effects of Head Start on children with differing incoming characteristics, particularly at lower levels of aggregation (e.g., by strata).

We conclude that use of the full battery for pretests on the total sample is the optimal course, unless fiscal constraints prohibit it. Any decision to limit pretesting either by reducing the number of tests in the pretest battery or reducing the number of children being pretested should await the results of the pilot study during the preparatory year on (1) the feasibility of random assignment in a sizable number of sites, (2) the costs of administering the basic battery, and (3) the analyses of pretest-posttest data as to differential effects. If the choice is made to reduce the pretest battery, then we strongly recommend that the Ravens Colored Progressive Matrices, CIRCUS Nos. 1, 2, 9, and 13 be included at a minimum, along with the CIRCUS Inventory of Test Taking Behavior.

Longitudinal Study

The second design issue regarding the battery recommended in this chapter is whether to invest in a longitudinal follow-up study to establish whether Head Start has any long-term effects in the perceptual-motor, cognitive, and language areas. We have addressed the general arguments against a longitudinal study in Chapter 1 and consider them applicable to the perceptual-motor, cognitive and language domains. There are two additional arguments against a longitudinal study which are specific to this domain.

First, the requirement that Head Start *should* accomplish cognitive gains beyond the year the child participates would make it unique among educational interventions. As Blank noted,[1] at most other levels the goal of schooling is to transmit information and skills in circumscribed areas (arithmetic, cultural history, French, etc.) and the expectation is that the student will be prepared to cope with the level of instruction immediately following. Poor seventh-grade algebra performance is not an indictment of third-grade arithmetic instruction, let alone of all of third grade, though it may be an indictment of sixth-grade mathematics instruction or the cumulative inadequacy of all prior math instruction. Nor is there any theoretical reason or experimental evidence to suppose that the limited Head Start intervention *could* meet the goal of raising the child's academic achievement at later educational stages. Thus, we reject a formulation of Head Start that requires the program to be judged by how long its effects last in the face of all kinds of intervening influences on the child.

Second, there is a considerable body of evidence already available on the long-term effects of preschool programs on cognitive and language skills. In general, some sustained gains have been documented in cognitive and language skills under two conditions: a preschool program that is specifically designed and controlled to achieve performance gains, and continuity of intervention across preschool and primary grades (see Ryan, 1974). As Bronfenbrenner (1974, p. 15) says, "the substantial gains achieved in the first year of group intervention programs tend to wash out once the program is discontinued."[2]

The Head Start program is not constituted to carry out these two conditions; in particular, it has no control over the child's experiences in the primary grades. Thus, it seems highly questionable to invest funds in yet another longitudinal follow-up of the effects of preschool on cognitive and language skills, when the results are predictable with some degree of certainty.

---

[1] Response paper by Marion Blank (1973) to Rand Cognitive Development Panel.

[2] Although this statement refers to IQ gains, it is applicable to our discussion because of the close correlation for older children between IQ score and intellectual achievement.

If the OCD deems it desirable to proceed with a longitudinal follow-up of the whole sample in the perceptual-motor, cognitive, and language areas, then what are the appropriate measures to be administered in the primary grades? For children whose first public school year is kindergarten, the battery proposed in this chapter could be readministered. For first grade and thereafter, there are three choices beyond teacher impressions and records (discussed in Chapter 5): (1) school grades, (2) standardized achievement tests administered by the school as part of its own testing program, or (3) a set of standardized tests chosen specifically for the Head Start evaluation and to be given to all children in the sample, regardless of local school testing. As far as being a valid follow-up measure to the Head Start test battery, there is no *a priori* reason to prefer one of these approaches over another. But for reasons of minimizing both the cost and the intrusiveness of the follow-up assessment, our preference of measures would be in the order of the above listing.

Although we are not sanguine about a follow-up study of the whole sample in this domain, there is good reason to consider some small-scale studies that would focus on linkages between Head Start and school. At least two different types of studies would be appropriate.

- Examination of the performance of children who go from Head Start to an enriched school experience, contrasted to that of Head Start children entering "regular" first grade. The study could be continued into succeeding grades, provided that enrichment experiences were available in these grades.
- Examination of match/mismatch between Head Start and school. One dimension for match/mismatch relevant to cognitive and language development might be the balance between open and highly structured learning situations in the Head Start classroom and the first grade that the Head Start child enters. Obviously, the match/mismatch question also applies to social and personal development, although studies aimed at these areas would have to specify different process and outcome variables.

Chapter 5

SOCIAL AND PERSONAL DEVELOPMENT

THEORETICAL OVERVIEW 153

ACTION SYSTEMS: ROLE BEHAVIORS TOWARD SIGNIFICANT OTHERS AND THEIR RESPONSES 158
  Role Behaviors Toward Peers and Their Responses 160
    Peer Evaluation 160
    Peer Interaction Styles 164
    Unobtrusive Peer Measures 167
  Role Behaviors Toward Teachers and Their Responses 169
    Teacher Evaluation 169
    Teacher-Generated Constructs 170
    Teachers' Evaluations Based on Standardized Constructs 171
    Teachers' Summary Estimates 173
    Interpretation of Evaluative Constructs 174
    Child-Teacher Interaction Styles 176
  Role Behaviors as Evaluated by Parents and Others 178
    Parent-Generated Constructs 178
    Parents' Summary Estimates 180
    Parent Involvement 181
    Interpretation of Evaluative Constructs 182
    Observers' Evaluations Based on Standardized Constructs 182

CHARACTERISTICS OF ACTION SYSTEMS: ACADEMIC 184
  Child-Task Interaction Styles 185
    Executive Skills 186
    Test-Taking Behavior 188
    Institutional Indices of Success and Failure 189
    Archival Data 190
    Scales of Early Adjustment 191
    Measure of Social Effects 192
  Learning Styles 193
    Direction-Following and Task Completion 193
    Goal-Setting and Self-Evaluation 197
    Intentional-Incidental Learning Cues and Reinforcement Style 199
    Curiosity 202

CHARACTERISTICS OF ACTION SYSTEMS: SOCIO-INSTITUTIONAL 205
  Role-Taking 206
    Spatial Perspective 207
    Situational Perspective 208
    Cultural Perspective 211
  Response Range 214
    Response Range to Nonpersonal Stimuli 215
    Response Range Relative to Interpersonal Stimuli 219

ATTITUDINAL CONSTRUCTS 223
  School Attitudes 224
  Self Attitudes 229
  Multiple Role Integration 231

Chapter 5

SOCIAL AND PERSONAL DEVELOPMENT

THEORETICAL OVERVIEW

This study's approach to conceptualizing and measuring social competence embodies "a theoretical orientation but not a theory."[1] The orientation is role-theoretic, inquiring how Head Start aids the process whereby individuals learn to enact the various social roles necessary for effective participation in the relevant social environment, construed here chiefly as the public school system. Such an orientation does not derive from an established developmental theory of role taking, for developmental research in this area is nonexistent. The role-theoretic orientation nevertheless appears to be useful for examining the "value-laden nature of the task"[2] of evaluating social and personal development. Role theory allows for the use of an evaluative notion such as "competence," without implying that any specific set of behaviors or attributes are intrinsic, situationally independent requisites for being a social "good" individual. Instead, evaluative questions are cast in terms of how well the individual perceives and responds to the role demands of his position in the social ecology.

Following Inkeles (1966) and Brewster Smith (1968), this definition can be elaborated to mean an ability to attain and perform in three sets of statuses:

a. Specific statuses in the social ecology to which one might appropriately aspire, such as the role of successful student (with its potential for increasing adult role options);
b. More generic positions that the society normally assigns, such as the proprietary norms related to age, sex, social class, and the like;

---

[1] Phrase borrowed from the Blocks' description of their own work (1973).

[2] Term used by Anderson and Messick (1973).

c. Individual statuses that one can reasonably invent or negotiate for oneself within the above constraints (including individual differences in style of role enactment, integration of multiple roles, and the resulting self construct).

Every person eventually must operate within all three domains. *The acquisition of relevant performance capabilities is what is commonly thought of as socialization, and the end or product of that process is competence.*

Given the view of social competence just described, what is the rationale for attempting to influence the social competency of lower income and minority children? The foundation for answering this question is provided by the *OCD-Head Start Policy Manual* (1973), in its focus on children's "everyday effectiveness in dealing with the environment," their ability to "cope with later responsibilities in school and in life." Head Start intends to insure for lower SES children the option of successfully attaining and performing in positions in majority culture institutions, starting with the public school. Existing literature (e.g., Coleman et al., 1966) indicates such children are at a tremendous disadvantage--academic achievement and social status apparently go hand in hand.

Many of the behaviors, styles, and attitudes learned as appropriate (proprietary norms) for lower SES role enactment and successful within certain environments are not those associated with success in secondary institutions such as the public school. Thus, children socially ascribed a low status in the second domain above incur a disadvantage in relation to goal attainment within the first domain, which Head Start hopes to offset. "Remediation" here has just the sense of overcoming obstacles that socially ascribed roles pose to the achievement of desirable statuses. The Head Start child is seen not as "disadvantaged" but as a child at a disadvantage with respect to certain outcomes. The treatment therefore is seen as an attempt to provide a supportive environment for the elaboration of the third domain: The Head Start child has an opportunity to practice a variety of role behaviors that will help him get around constraints associated with his position in the social ecology so that successful status attainment within a majority culture context becomes a genuine alternative.

Characterizing social competence and Head Start influences on it in the manner suggested necessarily involves presuppositions about the nature of outcomes and measures appropriate for evaluating them (cf. Bikson, 1974c).

> Assumption 1: *The child's interpersonal behavior strategies, along with personal characteristics and self-conceptions, evolve in response to the social context and, most important, in relationship to the significant others with whom he regularly interacts.*

Given this framework, outcomes discussed below are not intended to be interpreted primarily as personal characteristics, but rather as indices of situationally learned response styles. This assumption, derived from the interdisciplinary base of role theory, is consistent with current functional models of psychological adjustment and disorder (Dohrenwend and Dohrenwend, 1969; Hauser, 1971) that emphasize situational specificity of behavior styles, construed as responses to events occurring in a given sociocultural context.

> Assumption 2: *Emphasis on the importance of the socio-cultural context in child development cannot help but focus attention on the fact that children from different social environments occupy different social positions and learn different roles.*

The viewpoint taken here assumes (Dohrenwend and Dohrenwend, 1969) that what is perceived as socially desirable in one culture may be neutrally or even negatively perceived in another. Outcomes discussed below are therefore to be regarded not as cross-culturally invariant symptoms of positive psychosocial functioning but rather as aspects of action systems appropriate to successfully occupying the position of student in the public school system and other majority culture positions.

> Assumption 3: *Motor, perceptual, and cognitive processes are presumed to be common across cultures, even though they are used in quite different role behaviors.*

A culturally relativistic and situation-specific orientation does not imply the absence of developmentally common phenomena among children of similar age groups. In other words, the role-theoretic viewpoint assumes that genotypic similarities reflective of regular growth and maturation processes underlie phenotypic differences related to distinct subcultural roles. The preceding two chapters have identified some of the physiological and cognitive processes presumed to give the child the basic developmental skills necessary for learning a variety of social roles.

Determination of the nature of appropriate outcomes concomitantly includes where and when to measure them. Clearly, since the outcome sought is the Head Start child's effective participation in the secondary institution, evaluation must be directed toward the adequacy of his school role enactment. Moreover, because behaviors are reflective of and responsive to social contexts, the effect of Head Start will be most visible during the first public school year, gradually becoming confounded with the influence of school experiences themselves on subsequent performance. Social-personal outcomes, then, are proposed for evaluation mainly during the first public school year and in the public school context. Measures taken before public school entry or after that first year are recommended only on a limited sample and primarily for the purpose of augmenting knowledge of social development, rather than for assessing Head Start effects. And measures collected from parents are regarded as supplemental--they are taken not for the purpose of evaluating Head Start influence on the child's role in the family but for determining how parents view the child's enactment of the student role.

A difficult issue to confront is deciding what those outcomes are. Outcome classes, as noted above, cannot simply be derived from a well-established model of social competence development. Role-theoretic literature (Ziller, 1971; Brewster Smith, 1968; Sarbin, 1964) provides a suggestive general conceptual base from which the following outcome classes have been developed.

a. Effects are sought first in ongoing action systems, because in interpersonal relations--particularly as they involve young children--actions are of central importance. Further, they are visible and affectable (as opposed to more elusive intentions, attitudes, or traits). Action systems are construed in role-theoretic terms: Of chief interest are role behaviors toward significant others in the relevant social environment (peers and teachers), and the responses of significant others to those behaviors (expectations and evaluations) also constitute an important part of the action system, bearing on the assessment of adequacy of role performance. Both behaviors and responses to them among reciprocal role incumbents have considerable face validity as outcomes.

b. Operationally definable characteristics of action systems are a second important outcome class. These characteristics have been divided into two subclasses: behavioral modes that have been intimately linked (empirically or theoretically) with cognitive goals; and perception and response styles more generally associated with successful social behavior. The former characteristics are seen as most relevant to evaluating Head Start effects on academic performance, but the latter have potentially important long-term interpersonal consequences.

c. Attitudes are given lowest priority as outcomes for several reasons (even though socioemotional variables are often assumed to be attitudinal variables). Attitudes, first of all, are notoriously difficult to measure reliably in children, paper-and-pencil surveys being all but useless for subjects in the age range of the present population (yet this is the prevalent mode of attitude measurement). Second, the link between an attitude and subsequent behavior is always problematic; it is much less certain than the link between behavior and attitude or between behavior and subsequent behavior. Finally, attempts at attitude measurement among the proposed age group have not been very successful. Thus only two classes of attitudes are recommended for measurement, attitudes toward school and attitudes toward self.

Within the outcome classes developed, relevant research has been critically reviewed. Final choices of variables and their measurement depend on: the significance of the outcome variable in relation to Head Start goals as characterized above; the likelihood of its showing treatment effects; and its potential for providing considerable advance over past assessment attempts (cf. Chapter 1). Many existing variables with established instrumentation (e.g., "delay of gratification," "achievement motivation") are omitted for failing to fulfill these conditions. Many variables whose measurement requires pilot work are recommended because they do fulfill these conditions; however, it is assumed that final inclusion of any such outcome in the evaluation would depend on results of pilot study. It should be noted that cost and practical difficulty of incorporating an assessment into the evaluation battery is not regarded as critical. Rather, the primary aim here is to work out a set of assessments that cohere with the theoretic groundwork just presented in the absence of workable models of social competence development; it is assumed that, given the explanatory context and the signficance of each outcome, final decisions including cost and practical difficulty as considerations should rest with OCD and the research contractor.

Succeeding portions of this chapter treat each of the outcome classes represented above, in the order given. Following a summary discussion of kinds of information included in each of these areas of investigation, a more detailed discussion of selected variables and their measurement is provided.

ACTION SYSTEMS: ROLE BEHAVIORS TOWARD SIGNIFICANT OTHERS AND THEIR RESPONSES

[illegible]e "significant others" with whom the Head Start child engages in reciprocal role interactions as a student are primarily teachers and peers. Sources differ as to which of these interactions provides the strongest influence on social role learning (Kohlberg, 1969), but it is universally agreed that both are important. In particular, it has frequently been hypothesized that peer relationships are more important to the socially disadvantaged than to the advantaged child (Ausubel and Ausubel, 1963; Proshansky and Newton, 1968; Dohrenwend and Dohrenwend, 1969). More generally, Piaget (1948) proposed that the peer group provides unique role-taking opportunities for the young child since only here is he on an equal status with other role incumbents. Similarly, Kohlberg (1969) contends that simple frequency of participation in a social group is the primary requisite for effective role taking. Peer interactions will be examined either as processes or as evaluative responses to those processes.

Teachers are the role occupants who, along with peers, constitute the most important members of the school social system for the child. Although the child has an opportunity to perform more role negotiation and thus exhibit a broader range of appropriate relational behavior with equal-status peers, a significant part of school social behavior involves learning the role constraints of the student-teacher relationship. Learning to cope in role-appropriate ways with the teacher and the school setting represents the first interaction with secondary institutions (Kohlberg, 1969) and thus affords the developmental practice ground for broader social participation. Besides the developmental importance of learning to deal effectively with social institutions and their representatives, the nature of the child's interaction with the teacher is itself vitally related to his academic progress as well as to his psychological functioning. While this claim holds for all children, it is thought to be especially relevant to the lower status child who is both more wary of adults in the teacher role and more needful of their approval (Sarason et al. 1960; Ross, 1966; Zigler and Butterfield, 1968).

Teachers and peers, then, occupy the most important positions in the action systems to be explored in a study of school children's social competence. The nature of role behaviors toward them, and their responses, has highest evaluative priority. In addition, two other sources of action system data might relevantly be investigated. One data source is the child's parents, who can be regarded as relevant role incumbents with the following reservations. Although parents have been the most significant adult others during most of the child's life, the parent-child relationship is both too broad and too narrow to examine in a study of social competence. It is too broad in that it encompasses much more than social competence; anyone can confirm this claim by considering differences between social-self and family-self factors in the self construct (e.g., Tennessee Self Concept Scale, Fitts, 1964). It is not at all clear that Head Start should be evaluated as an intervention in the child's family role, except insofar as his relationship to his family is just another instance of general social relations. For the purposes of this evaluation, however, relationships within the family are too narrow a base of study. Socialization is usually seen as a process, aided by training at home, during which the child becomes a member of the broader social system outside the home. School or preschool is usually the first opportunity a child has to participate fully in a group other than the family, and thus this setting is the area in which social competence effects of Head Start are properly sought. Hence, parents are regarded, for the purposes of this project, as relevant sources of information only as parents of students.

Test administrators and observers who have occasion to deal with subjects during the course of the evaluation research are "secondary institution" representatives (Kohlberg, 1969) related to the child only by virtue of role occupancy (i.e., the child does not know them personally). But part of developing social competence is being able to respond appropriately on such a basis. Such other perceptions of the child in the assessment situation, then, are a potential source of data for analyzing the range of the child's social role repertoire.

Role Behaviors Toward Peers and Their Responses

Peer Evaluation. In his review of research concerning psychosocial functioning in elementary school, Bower (1960) points out that almost all studies of peer perception point to a strong relationship between emotional adjustment and peer judgments. Concurring, Kohlberg, LaCrosse, and Ricks (1972) find stable, accepting peer relations to be important antecedents of later adjustment, peer evaluations having more predictive power in this regard than clinical ratings; their review suggests that the most useful socioemotional indices are extremes of peer acceptance or rejection as ascertained in repeated observations. Operational referents for peer status variables typically include both verbal and nonverbal responses which a child elicits in (relevant) other children. These terms commonly denote the degree to which a child's peers wish to engage in some form of associative contact with him (Kimbrough and Bikson, 1972). According to LaCrosse (1974), such relationships are rapidly stabilizing in kindergarten and first grade.

Methodological solutions to the measurement of social acceptance are commonly termed "sociometric techniques" and derive from the model popularized by Moreno (1934). While a variety of sociometric formats is available *for a large-scale study with the age range of the present target population, a simple self-report is recommended.* This version of the sociometric technique requires a situation in which the subject is individually presented with one or more criteria (unqualified friendship, work companions, outdoor play companions, etc.) and is asked to nominate from among his classmates a certain number of positive and negative candidates (most preferred and least preferred choices). The sociometric task has been administered in this form to kindergarten-aged and younger children of varying SES by Gerard and Miller (1971), Kimbrough and Bikson (1972), Boger and Knight (1969), Jensen and Kohlberg (1966), Cassel and Martin (1964), Bonney and Nicholson (1958), and several other researchers reviewed in Walker (1973).

From the work of Boger and Knight (1969) it seems evident that the use of five critical situations is too taxing for children in the projected age range. Gerard and Miller (1971) successfully elicited responses to three critical situations from kindergarteners, but requested only positive nominations. *Because it seems that negative nominations are as important indices of peer competence as positive ones* (Cowen et al., 1965), *and because this procedure doubles the number of choices per criterion, a single-criterion task is recommended here.* Moreover, for the age group in question, an undifferentiated friendship choice is probably most appropriate. Number of choices requested in sociometric tasks has varied from open-ended (Ziller et al., 1969) to eight at the upper limit with two as the lower limit. The experience of Jensen and Kohlberg (1966) with both ends of this range suggests that at best three choices can be successfully elicited, the number also successfully used by Cassel and Martin (1964) and by sources reviewed in Walker (1973). It is concluded, then, that *three positive and three negative sociometric choices should be elicited from each child.*

Although choices can be elicited with or without the help of visual aids, Boger and Knight (also McCandless and Marshall, and Moore and Updegraff, reviewed in Walker (1973)), think that a board on which a picture of every child in the class is mounted facilitates the choice-making procedure. Kimbrough and Bikson (1972) also recommend it as helpful in scoring, since children often refer to their playmates by nicknames, middle names, or other "unofficial" designations; being able to see who is meant, by having the child point, insures that nominations are properly recorded. Using the picture board, it is also possible to see how many classmates the child can, in fact, name. Boger and Knight (1969) regard this as a good warm-up before choices are actually made, while other sources elicit the remaining names afterward. In either case, both Boger and Knight (1969) and Jensen and Kohlberg (1966) think the number of classmates known (that is, nameable) by the child is itself a social variable worth recording. *The picture board procedure, with elicitation of classmates' names, is recommended for the present study.*

The recommended procedure, then, should yield the following kinds of information about peer evaluation:

1. Peer acceptance and rejection, measured as the number of positive and number of negative choices received.
    a. Socially underchosen children are those who receive fewer choices than they make, social isolates being regarded as those who receive none (or sometimes one or none).
    b. Percentage of negative choices to total choices received indicates peer antipathy, regarded as perhaps the single most sensitive sociometric measure (Cowen et al., 1965; Bower, 1960).
2. Social reality of friendship choices, measured by reciprocity of nominations.
3. Ethnic parameters of choices, represented by number of inter- and intra-racial nominations.
4. Other:
    a. General social aptitude, reflected in number of classmates whom the subject can name; if the measure is taken more than once during the school year, it would be expected that this measure would show a ceiling effect (i.e., it would only be useful as an indicator of rapidity of school social adjustment).
    b. Stability of sociometric status, reflected in constancy of the three measures above if collected more than once during the school year, would provide indices with greater longitudinal predictive value (Kohlberg, La Crosse, and Ricks, 1972).

The measures discussed above have been subjected to two sorts of reliability studies, those focused on the degree to which the same child makes the same nominations in the same order from one occasion to the next, and the degree to which the same child receives an approximately similar popularity rank from one occasion to the next. The former sort of study (Boger and Knight, 1969) indicates that children do not reliably produce the same choices in the same order, although there is considerable overlap when order is disregarded. For purposes of the present study, it is recommended that order of nomination not be regarded--i.e., that peer nominations not be weighted and that a child's sociometric status (positive or negative) be computed in terms of the simple number of nominations received. When order is disregarded, reliability coefficients are acceptable, ranging from 0.45 when three criteria are used to 0.86 with a single-criterion test (reported in Walker et al., 1973). Although no developmental norms are available for sociometric measures, their validity is generally acceptable (Walker, et al., 1973); Bower, 1960; Jensen and Kohlberg 1966).

It is recommended that the peer nomination measure be taken at least once during the first month of the first public school year along with other measures of ease of adjustment. It would be desirable to repeat the measure at least once later during the school year to determine the stability of the relationships initially obtained and the persistence of Head Start effects (if any). The choices-received measure in a single-criterion sociometric task (Bonney and Nicholson, 1958) has been found to discriminate kindergarten through third-grade children on the basis of preschool experience. Because repeated administrations would not need to include the naming of all children in the classroom, the task should be very brief and easily incorporated in any other later data collection sessions. It would additionally be interesting to include this measure during the Head Start year, although no comparable control group measure would be available. Such a procedure would provide information about how a child's sociometric status changes as he moves from Head Start to public school; that is, it is conceivable that a child be in the top sociometric quartile of his Head Start group and move to the bottom sociometric quartile of his public school class even while retaining a sociometric advantage over control group children equivalent in SES. Such a result might well occur if the general SES level of the classroom is higher than that of Head Start-eligible children, and might be useful in interpreting data related to school attitudes and self attitudes.

Peer Interaction Styles. Peer interactions are the behavioral processes of which sociometric nominations are presumably the evaluative outcomes. Thus all that has been said concerning the importance of peer relationships for inferring social competence holds directly for the study of peer interaction styles. Bronfenbrenner (1969), Zigler (1970), Butler (1970), and others have underscored the significance of social functioning in the peer group for the formation of social values, yet this process has not been widely studied. Both Stearns (1971) and Kohlberg, LaCrosse, and Ricks (1972) have pointed out that the study of intra-individual traits in children has provided little concrete understanding of the role behaviors involved (e.g., what behaviors are included in the notion of a self-reliant or cooperative child) and no basis at all for predicting later social adjustment. In contrast, while overt behaviors are much more accessible than traits, developmental research seems to have ignored the former in favor of the latter (Stearns, 1971; Zigler, 1970). It is, consequently, desirable to have an empirical study of the role behaviors that are associated with peer sociometric status (as well as with other social outcome variables discussed below). It is additionally desirable to determine what sorts of prosocial behaviors can be increased by preschool experience and what sorts of negative behaviors can be decreased by such experience, assuming (see Anderson and Messick, 1973) that these two broad event classes are somewhat independent. Further, as White (1973) points out, it seems likely on the basis of existing evidence that *some* preschool programs do increase *some* prosocial behaviors, but because socioaffective measures are too global and of uncertain validity, it is difficult to be confident about just what these effects are. Measuring ongoing action systems should circumvent this difficulty and yield valuable social competency information.

The conclusion reached by White (1973) as well as by Walker (1973), Walker et al. (1973), and Stearns (1971) is that for young children the most promising and most effective socioemotional measurement strategy is the structured observation technique. After an exhaustive review of existing measures, Walker (1973) argues that observation techniques for assessing social competence avoid the major difficulties of other measures (social desirability and other response-style sources of bias, dependence on cognitive ability, and cultural specificity) by being nonverbal and objective. Another substantial advantage of observation methods is their ecological validity. Available literature indicates that the bulk of contemporary child development theory rests largely on an empirical base generated either in the laboratory or in clinical case studies (Kimbrough and Bikson, 1972), and its application in the natural field setting is tenuous. Thus, *structured behavioral observations in the natural setting are necessary to ensure that the social competence evaluation does in fact represent the child's "everyday effectiveness in dealing with the environment"* (OCD, 1973).

Methodological procedures for structured observations are fairly well established, stemming largely from studies of attempts to modify the behavior of children in the classroom (e.g., Bijou and Peterson, 1969; Baer, Wolf, and Risley, 1969; and numerous recent studies in the *Journal of Applied Behavior Analysis*). For the proposed evaluation research, pilot work will be required to prepare a structured observation instrument for a large-scale field assessment of social competence with peers. However, such piloting is needed primarily for selecting among techniques and target behaviors represented in the sources cited here, which constitute a solid body of groundwork. The most important decisions to make regarding structured observation of peer interactions concern selection of target behaviors, which is expected to be the most important part of pilot work. Such work should be guided by the role-theoretic approach outlined above, an approach congruent with the definition of social skill offered by Bronson (1973) as a preface to her behavior coding system:

> SOCIAL SKILL is defined as the ability to control and direct oneself adequately and constructively in social situations and the ability to influence others effectively in socially approved ways. Since approved methods of social control of others vary with the culture or sub-culture, any assessment of a child's competence in this area necessarily implies value judgments. The judgment implicit in the categories of this profile is that a general attitude of negotiation and reciprocity in dealings with others is a desirable goal.... This implies the ability to control or influence others with effective but non-violating strategies (physical force is considered to be a strategy which violates the social other and therefore does not show an attitude of negotiation and reciprocity), and the balancing ability to be reasonably influenced by the group without being totally overcome or dominated by others.... It is assumed to reveal an awareness of general rules for social interaction which apply across specific social situations and independently of specific individual wishes--e.g., social contract rules. Specific strategies which facilitate social interaction and promote cooperation such as sharing, helping and combining resources are especially noted.

Target behaviors, then, are empirically determinable role-appropriate and role-inappropriate interaction styles with peers in a public school setting. Ogilvie and Shapiro (1972) believe it is not necessary to code all aspects of social behavior, but only those having most relevance to the notion of social competence. The description of procedures whereby Ogilvie and Shapiro derived their code categories (in Walker, 1973) lends considerable support to their construct validity as a whole. Moreover, their categories include representations of most of the following outcome areas deemed important on the basis of a fairly thorough review of preschool peer-behavior literature: (1) degree of social integration during free play periods; (2) affective valence of contact; (3) ascendancy-submission; (4) modality of contact; (5) purpose of contact; (6) ethnic aspects; and (7) stability of social interaction style.

It is recommended that the Ogilvie and Shapiro categories form the basis for pilot work on a structured observation coding system for peer interaction styles, perhaps as supplemented by categories represented in Bronson indicating more fully the extent of social participation involved in a given peer contact. It is further recommended that ethnic aspects of peer interaction be included in the coding scheme. Finally, stability of interaction style can be assessed only if observation data are collected more than once during the school year. While it would be desirable to have at least two such data points for the entire sample, it is admittedly a costly and time-consuming procedure. Because repeated observation on the entire sample is not feasible, it is recommended that after initial whole-sample observation, repeated observations be made on a subsample.

Whatever final form the proposed observation system may take, reliability should not be difficult to establish. For the Ogilvie and Shapiro system, an overall reliability coefficient of 0.87 was computed on paired 1/2-hour observations of 20 children aged 3 to 6 in seven preschools (Ogilvie and Shapiro, 1972). For most observation instruments, norms are not available. However, scores on eight social competency dimensions from the Ogilvie and Shapiro behavior coding system are available for children aged 12 months to 33 months, a population somewhat below the age range of interest for the present evaluation project (reported in Walker, 1973). Construct validity of code items in that observation instrument is evidenced by their ability to discriminate highly competent and noncompetent preschool children (see Walker, 1973). Further, two summary items used in the SDC study (1972), total verbal interaction scores and total frequency of initiations of interracial contact, yielded significant pretest to posttest changes during the Head Start year. SDC interpreted these gains to mean that Head Start makes an important positive contribution in the socioemotional domain, implying that such measures might well distinguish Head Start children from the control group with respect to social competence. For any observation system, however, face validity is perhaps as important as any other consideration. That is, outcomes in this area are directly observable behaviors regarded as desirable by the social system in which the child aspires to a socially valued status, and they do not need to be validated by their relationship to anything else.

Unobtrusive Peer Measures. Peer nominations are too global to capture many aspects of classroom interaction, and observation instruments are too costly and time-consuming to be used frequently, but a compromise is perhaps possible. Environmental studies of children's behavior have attempted to determine what play equipment is most frequently used during free play by mounting "fish eye" cameras high enough and at appropriate angles to include the entire play behavior setting. The cameras are then connected with electronic timing devices, which, when activated, take still photographs at a predetermined rate (e.g., one per minute). The photographs are later scored with respect to relevant play parameters. Electronically timed photographs could be used to determine on a time-sampling basis the frequency of isolate, dyadic, or n-adic contact among children in the free play situation by scorers wholly unrelated to the classroom situation; by scorers familiar with the children, confirmation of sociometric and observation data could be obtained by recording who was associated with whom in each photograph (perhaps noting cross-race and cross-sex play contacts and group size as well). Still photographs are preferred over videotapes because of the automatic data reduction involved and greater ease of scoring. Such a measure of peer interaction is technologically feasible and potentially less costly than most process measures. *However, any use of this photographic sampling technique at present would be wholly exploratory and recommended only for a small subsample study.* (See Walker, 1973, on the need to fill the instrumentation gap between sociometric solutions and observational procedures for indexing development of peer interaction styles.) It therefore has lowest priority among peer measures.

In summary, then, the following recommendations for assessment of subjects' role behaviors among their peers have been made. Interpersonal peer processes are to be assessed by structured observation of interactions as they occur during free play periods. Such observation data should be collected once near the end of the first month in public school for the entire sample, repeated observations being collected on a subsample basis. Evaluative responses to subjects' interpersonal behavior styles will be obtained from their classmates by means of sociometric nominations to positive and negative status categories. Like the process data, sociometric information should be collected once near the end of the first school month for the entire sample and again when repeated process measures are taken for the same subsample. Finally, a subsample peer-interaction study is recommended using photographic time-sampling of the behavior setting.

Role Behaviors Toward Teachers and Their Responses

Teacher Evaluation. As with the study of peer relationships, teacher-child interactions can be examined either as processes or as evaluative outcomes of those processes. Both sorts of studies are recommended here. The importance of teachers' evaluative responses to children in the classroom cannot be underestimated, as the following brief review will indicate. First, it has frequently been found that lower-status children, especially minority group members, experience the classroom situation and confrontations with the teacher as occasions of social threat and failure threat; that is, this social setting is perceived as threatening and hostile, and it elicits strong expectations of negative evaluation (Katz, 1964, 1968; Sarason et al. 1960; Proshansky and Newton, 1968). Second, evidence gathered in interview situations with low-SES children (Williams, 1970a; Williams and Naremore, 1969; Labov, 1970) suggests that their reaction is one of caution, adopting the least threatening strategy and making minimal commitments to the teacher *qua* teacher. The child's behavior then tends to confirm the teacher's *a priori* expectations that low-status and minority children are "passive," "not trusting," "not spontaneous," "not sociable," and so on (Williams, 1970a; Gerard and Miller, 1971; Bikson, 1974b). Third, teacher expectations are potentially self-fulfilling (see the literature reviewed on this topic by Rosenthal and Jacobson, 1968), both in positive directions (Rosenthal and Jacobson, 1966) and in negative directions (Gerard and Miller, 1971). Although most sources have hypothesized that children's performance is mediated in significant ways by attitudes that arise in the course of teacher-child interations, very few studies have probed this process, looking rather at more removed outcomes such as achievement scores and "self-concept" measures. *A study is recommended of teachers' evaluative responses to children early in the school year along with a study of the actual behavioral styles with which they are associated.* Head Start could make a strong contribution to the lower-status child's classroom success if it succeeds in changing the child's approach to the school setting in such a way that the dynamics of teacher-child interactions result in fewer self-fulfilling negative expectations during the school year.

Teacher-Generated Constructs. Two sorts of teacher evaluations are recommended, each based on well-standardized procedures. The first elicits teacher evaluations based on the teacher's own set of role expectations and is based on Kelly's (1955) role-construct repertory test. It has been pointed out that there may well be regional differences in teacher expectations regarding student behavior, as well as ethnic differences and differences based on other factors such as number of years of teaching experience (SDC, 1972; Coleman et al., 1966; Gerard and Miller, 1971). The most useful feature of the Kelly test is that it allows the individual teacher-evaluator to generate the constructs on which the subsequent evaluation is to be based; the extent to which children are satisfactorily fulfilling their own teacher's role expectations--whatever they may be--can then be determined.

It has been suggested that perhaps the safest meaning to give to the notion of social competence is perceiving and responding in acceptable ways to the role-standards of the immediate social community.[1] The repertory test establishes student role standards in the social community of which public school teachers are representatives by asking teachers to focus on examples of very good and bad students in their recent experience, generating a concrete frame of reference. Characterizations provided by teachers are then translated into rating scales by which current students are judged. While different teachers will generate different evaluative dimensions, it can be determined whether there are differences between Head Start and control children in overall favorableness of evaluation. Further coding of teachers' characterizations should allow more refined comparisons. It should also permit examination of differences among teacher groups regarding student role standards. *These evaluations should be collected once for the entire sample during the first month of school, and should be repeated for a*

---

[1] This suggestion was made independently by Dr. Gloria Powell (child psychiatrist, Martin Luther King Hospital, Los Angeles) and Prof. Millard Madsen (cross-cultural developmental psychologist, UCLA).

subsample. In addition, it might be worthwhile to administer this sort of rating scale to a subsample of Head Start teachers to determine the degree of congruence about student role expectations among those teachers and public school teachers in the same area as well as the degree of evaluative agreement regarding students. Control children could not, however, be subject to the same pre-measure.

A full discussion of the general reliability and validity of the role-construct repertory test is available in Kelley (1955). It should be pointed out, however, that the scales cannot be presumed to give objective information about a child and cannot even be examined for interrater reliability since they are teacher-generated. What they can be presumed to provide is a good guide to how the teacher articulates student role-qualities to herself and how well the child has perceived and satisfied these expectations. Thus, the scales provide important information about the teacher-child interaction.

Teachers' Evaluations Based on Standardized Constructs. In addition to the sort of teacher evaluation just discussed, it is also necessary to determine how Head Start children are faring compared with control children with respect to generally accepted evaluative dimensions. Such judgment dimensions would be taken to represent a social consensus regarding the constructs most important to student role performance. Teacher ratings have been used in most preschool and elementary school evaluation studies reviewed (Jensen and Kohlberg, 1966; SDC, 1972; Walker et al., 1973; Anderson, 1960; Wolff and Stein, 1967; Ward, 1973). While some sources express concern about teacher bias (Stearns, 1971; Walker, 1973), the viewpoint taken here is that the teacher is a central role-incumbent in the classroom social setting. Thus, although the teachers' judgments are probably not indicative of objective and cross-situational child outcomes, they do represent something consequential about the current position of Head Start children in the eyes of the institution in which it is hoped they will come to achieve a socially valued status.

There is no shortage of teacher rating scales or standardized procedures for using them (Walker, 1973). The only difficulty concerns which among the many rating instruments is most suitable for the present purpose. A review of the literature provides validation of three major factors around which teacher judgments tend to be based; the first higher-order factor represents a "love-hositility" dimension; the second, an "introversion-extroversion" dimension; and the third, a "task-versus-person orientation." That is, diverse investigators have started with fairly large nonidentical item pools (ranging from 72 to 200), and have used teachers or independent observers to rate large groups of children on each of the items (using scales of varying refinement). From first-order and second-order cluster or factor analyses, this basic tri-component response structure continues to emerge (Baumrind, 1971, 1972, 1973; Baumrind and Black, 1967; Becker and Krug, 1964; Emmerich, 1971, 1973; Schaefer, 1961; Kohn and Rossman, 1972; Walker et al., 1973). That diverse--and even uncongenial--streams of research converge so consistently on these distinguishable aspects of children's classroom competence provides convincing grounds for recommending the use of rating scales based on them.

It is recommended here that a small item-pool sampling the content of the three second-order factors be used as the teacher rating instrument. Such a short form is represented by the Classroom Behavior Inventory (Walker, 1973; Walker et al., 1973). For this 15-item inventory there are norms based on the total fall 1971 HSPV sample, a large and ethnically diverse subject group. Reliability studies (Walker, 1973) suggest that test-retest reliability is sufficiently high (0.70's) but that inter-rater reliability is not as high as it should be (ranging from medians of 0.62 and 0.60 on task-orientation scales to medians of 0.39 and 0.44 on hostility scales). Most of the difficulty is explained by rater style: Some raters tend always to use extremes on scales while others tend always to stay toward the middle; and social desirability biasing produces some tendency toward a ceiling effect on task orientation items and toward a floor effect on hostility items. This circumstance also makes scores nonaggregatable, which severely impairs the usefulness of the inventory in its present form for a national study. To alleviate distribution difficulties and allow for data aggregation, it is recommended that children be rated on these 15 7-point scales using a constrained Q-sort. Such a rating procedure should also be followed in obtaining responses to the evaluative dimensions teachers have generated themselves, so that these two teacher instruments will yield comparable data sets. It should be noted that rating on standardized scales must be done *after* role construct rating; otherwise constructs elicited from teachers will have been influenced by the content of the standardized instrument. However, it could be done immediately afterward, during the same session.

Teachers' Summary Estimates. Teacher summary evaluations should be sought on the four points listed below, which have some face validity but are primarily useful as unobtrusive indices of teachers' attitudes toward sample children. These evaluations may be included at the end of the preceding rating session or else may be obtained any time near the time when observation data are collected. Very brief piloting should be undertaken to determine the easiest sort of scaling: asking teachers to make decile estimates (e.g., the top 10 percent) produces a good range of possible values, but making estimates in the middle deciles may prove difficult; in contrast, Dunnington (cited in Walker, 1973) asks teachers to place children in top, middle, and bottom thirds, which yields little variation in scores. In any case, *it is necessary to locate the child in relation to his classmates*. The four areas for summary evaluation are:

1. Estimate of the child's sociometric status with his peers. This judgment will enable determination of the extent to which teacher ratings reflect peer ratings; usually the correlation is not high. It is recommended that the direction and distance of deviation of the teacher's judgment from the peer rating be regarded as a measure of the teacher's evaluation of the child's peer interaction style.
2. Estimate of the child's "school adjustment." This estimate is sometimes used to compose a criterion variable against which socioemotional measures are judged for validity (Lambert, 1963). It is expected to correlate with the previous judgment and provide a general social indicator variable.

3. Estimate of the child's upcoming or immediately prior achievement test placement. It is expected that all children in public school will either have taken some form of "readiness" test, will be about to take one, or will have to take one at year's end. The response provides an index of the teacher's academic expectancy for the child, which should be guided by an "accuracy" orientation--i.e., the teacher will want to guess as correctly as possible. Thus the estimate should not only indicate real expectancy but the direction and distance of deviation from accuracy should again provide a measure of the extent to which the child *qua* student has made a favorable impression on the teacher. While teachers have been found to have unduly low expectations for lower-status minority children (Gerard and Miller, 1971), any differences between control and Head Start children on this variable would indicate that the child's performance of the student role was beginning to change teacher expectancies.
4. Estimate of motivation. This is another summary evaluation of the child in his student role. It asks whether the teacher perceives the child as making a real effort to fulfill student role demands and is expected to correlate with the immediately preceding estimate; if the child is further perceived as trying harder than the teacher thinks his test scores will indicate, the child has clearly established himself as a student.

Interpretation of Evaluative Constructs. This discussion attempts to point out an area where some exploratory work needs to be done to determine whether the same evaluative terms have the same meaning when applied by teachers to children of different status groups. While the problem does not appear among the conceptual difficulties elaborated by Anderson and Messick (1973), it seems to be one of the thorniest. It is entirely possible that the terms "assertive" and "independent" might be applied with a positive connotation to white middle-class children and with a negative connotation to Black lower-status children; it seems likely that "withdrawn" might be a characterization that would invoke sympathy if applied to a young white child while invoking suspicions of passive aggression if applied to a young Black child. Perhaps these fears are unfounded, but the study by McNeil and Phillips (1969) with sixth-grade students established that traits positively related to school success among white children were not so correlated in Black and Chicano children. Other studies have found that terms regarded as "complimentary" to Blacks by whites were not so regarded by Blacks, and conversely.

Exploration of this question might take two forms. *First, it would be fairly unproblematic to administer one of the large item-pool rating instruments (from which the Classroom Behavior Inventory above is drawn) to a subsample of teachers*, making certain that a large representative sample of each major status group of school children was rated (lower-income Black, Chicano, and Caucasian children, as well as middle-to-upper-status white children, at minimum). After such ratings, factor analyses should be undertaken to see whether the same basic factor structures are replicated within each status group and what high-loading items defined each factor. To our knowledge, this has not been done, yet replicability of the factor structure within status groups is necessary if the results of the ratings are to be interpreted sensibly. Alternatively, a multiple discriminant analysis might be performed on overall factor scores to see whether, within a single response structure, functions differentiating subcultural groups result.

A second exploratory effort might make use of semantic differential techniques. These techniques are themselves well established (Osgood, Suci, and Tannenbaum, 1958). They would be used here to determine the extent to which the same adjectives loaded in the same way on the same factors when used to characterize children of different ethnic groups. For this purpose, a large subsample of teachers could be used as raters, and a fairly lengthy list of bipolar adjectives could be devised. What to use as stimuli for eliciting teacher responses is the main question to be answered. In this case, no specific individual children are evaluated; rather, an attempt is made to indicate connotative meanings inherent in teachers' conceptions of children of given status groups. For this purpose, picture stimuli (e.g., photographs of one or a small group of children representing each of the status groups being studied) or descriptive phrases (e.g., "Head Start child," "middle-class Black child," "lower-income Chicano child") or perhaps brief taped messages might be used (e.g., "My name is __________; I go to __________ school; I'm in the first grade and my teacher is Ms. __________," where fictitious names were used). While semantic differential stimuli are typically words or phrases, they seem here to be too blunt and might result in considerable social desirability response biasing. Pictures have, however, been used successfully as stimuli in semantic differential studies, and it is also clear that teachers pick up ethnic cues very quickly from brief taped speech samples (Bikson, 1974b). Various forms of stimuli could be piloted, and, whatever the final choice, not all teachers would have to rate all stimuli. Factor analysis within status groups would indicate the extent to which the same adjectives loaded positively on the evaluative factor for each. Replicability of factor structures would indicate that the same terms had the same evaluative connotations when applied to diverse status groups, lending more confidence in the interpretation of teacher evaluation data. Because these latter studies do not contribute assessment data but only aid in their interpretation, they have lowest priority in the set of teacher responses collected.

Child-Teacher Interaction Styles. As indicated previously with respect to peer relationships, children's interaction styles with teachers are taken to be the processes of which the evaluations are results. It should be clear from the foregoing discussion that, even though teachers are able to give extensive verbal evaluations (unlike the child's peers), it is no less important here to be able to associate those evaluative responses with specific classes of behaviors. Further, the behaviors that are properly regarded as "prosocial" toward peers might be regarded as inappropriate toward a teacher, who occupies a quite different status. The notion of social competence, however, entails being able to make such distinctions. Finally, it has been suggested on the basis of rating scales (Emmerich, 1971; 1973) that actions toward teachers are more closely related to the child's school task orientation than to a person orientation. Thus, there is good reason to recommend that structured observations be directed specifically at child-teacher interactions.

Fundamental procedures involved in devising a structured observation instrument for teacher-child interactions and for child-child interactions are similar. As before, it is recommended that nonprofessionals be trained in the reliable use of a code that samples behavioral events, and that the sampling be done during approximately the same time as the collection of teacher evaluation data. In particular, it seems advisable to have the same observers perform both tasks, the peer interactions to be scored during a free-play period and the teacher interactions to be scored during an indoor semistructured period. It is recommended that some sort of art or craft period be observed.

Clearly, pilot work is needed to finalize technical aspects of behavior scoring as well as to determine the target behaviors to be scored. With the exception of ethnic aspects (which are not subject to voluntary variation during teacher-child interactions), the outcome areas listed for peer relationships apply in the study of teacher-child relations. *It is again recommended that the pilot study take the Ogilvie and Shapiro (1972) teacher-child interaction categories as a starting point in determining target behaviors for these outcome areas.* It would be desirable to supplement the Ogilvie and Shapiro categories so that they indicate more fully the extent of social interaction with the teacher, specifically as suggested by the code categories in Grotberg (1969) and by the categories used by the U.S. Commission on Civil Rights (1973) in its study of Chicano children in public schools. These latter systems code what the teacher does with a child's input. (In contrast, Ogilvie and Shapiro record whether attention-getting, help-getting, etc. were successfully or unsuccessfully attempted, but they cannot record a more extended interaction in which the child makes a remark and the teacher elaborates on it or otherwise continues the interchange.) Grotberg also makes provisions for recording the child's reaction to the teacher's response, which seems an important indicator of how the child is coming to terms with institutional authority figures. Finally, the definitions of efforts to control an adult and to comply with or resist complying with adult direction in Ogilvie and Shapiro should be expanded to include more concrete examples of appropriate versus unrealistic limits-testing with a teacher.

Questions of reliability, norms, and validity of this sort of structured observation instrument have been previously discussed. Two points are worth restating. First, observational data are selective in that they can represent only a part of what goes on in the classroom; but that limitation notwithstanding, they are the best source of objective information on the basis of which judgments about competence can be made. Second, because the data are observational rather than inferential, they provide concrete referents for what a teacher means by "sociable" rather than "demanding," "independent" rather than "defiant," and so on. Thus observations of child-teacher interactions provide an objective account of classroom interaction styles, a basis for interpreting teachers' evaluative judgments, and a means of determining the responsiveness of the latter to variations in the former.

In summary, the following recommendations have been made for assessing subjects' role behaviors in relation to their teachers. Teacher-child interaction processes are to be assessed by means of an event-sampling observation instrument focused at semi-structured tasks. Teachers' evaluative responses to those behaviors will be investigated using three sorts of behavior rating methods: (1) teacher-generated evaluative constructs; (2) standardized evaluative constructs; and (3) summary performance estimates. In addition, exploratory investigations of the meaning of evaluative dimensions are suggested. Process data should be collected along with the first three sorts of ratings near the end of the first month of public school for the entire sample and should be repeated for a subsample. The repertory test (rating method (1) above) was suggested as a pre-measure for the Head Start teachers of a subsample of treatment subjects.

Role Behaviors as Evaluated by Parents and Others

Parent-Generated Constructs. The notion of "competence" implies living up to some standard(s). "Social" competence, from the standpoint of cross-cultural developmental psychology, can mean nothing other than living up to the standard(s) set by one's community, relative to one's social role.[1] By this definition, the Head Start child is socially competent to the extent that he satisfies the role expectations held for him as a school child by the adults who care for him.

For the purpose of determining what parents' role expectations are for the child and how well the child is satisfying them, a version of the Kelley (1955) role-construct repertory test is recommended. While some pilot work would have to be undertaken to revise instructions and procedures for obtaining parent evaluations, basic steps would be similar to those suggested for teachers. That is, parents would be asked to consult their own previous experience, focusing on instances of exceptionally good and bad school children (e.g., older siblings of the target child as well as children of friends and neighbors, or even their own former school-mates). These cases then serve as familiar, concrete anchor points for generating evaluative construct dimensions. Once evaluation dimensions are established, the parent performs a standard behavior rating task in which his or her child is the subject rated. Results would be useful in many ways. First, data so obtained would indicate what role standards parents have for their children as students and how well children are living up to these expectations. Evaluative scores can be compared to determine whether Head Start and control children differ in this regard. Second, it would be equally interesting and important to see whether Head Start parents have expectations about student role performance that are more similar to teachers' expectations than are those of control parents; significant differences in this direction might be expected and, if confirmed, would indicate the influence of Head Start on family as well as child social characteristics. Finally, degree of similarity and discrepancy regarding role standards among parents of different regional and ethnic backgrounds can be examined, with a view toward determining differential deviation from modal teacher expectations. Such information should be valuable for estimating difficulty of role adjustment among different subject groups. It is recommended that this information be obtained on the same schedule as teacher evaluations--[illegible] once in the fall for the entire subject sample, with repeated measures taken on a subsample. Pre-measures should also be obtained for a subsample, in order to observe changes in parent expectations resulting from their relationship to the Head Start program.

---

[1] Suggested by Prof. Millard Madsen, psychology department, UCLA.

Questions about reliability and validity of the Kelley technique have been discussed above. One further point should be noted here. The difficulty of interviewing parents of low-SES children using standardized survey instruments has been attributed in part to the medium (the language of test construction is unfamiliar) and in part to the message (the content is culture-biased). The role-construct test, in allowing the respondent to supply item content in his own terminology, helps circumvent these difficulties. Such difficulties do, however, argue against the use of standardized behavior rating scales among parents.

Parents' Summary Estimates. *Parent summary estimates will be used in part as face valid indices and in part as corroborative data for conclusions drawn from other sources regarding the child's school attitudes.* Piloting should determine, as with teacher summary estimates, the preferred manner of scaling; it would be hoped that similar scale intervals could be used, to facilitate comparisons. Here intervals cannot represent ipsitive ordering, since the parent does not know the other children in the class. Thus intervals will have to be given some other ordinal scale interpretation for parents.

1. Estimate of the child's sociometric status with his peers: This judgment will provide a reflection of the child's interpretation of his social standing in the class as he conveys it to his parent. Of interest, beyond the differences in parent inferences concerning social standing of Head Start versus control children, are comparability of parent estimates both with teacher estimates and with real sociometric rank.
2. Estimates of the child's school adjustment should be considerably more detailed than those obtained from the teacher, focusing on:

a. The child's overall happiness with the school situation;

b. The child's reluctance to leave home for school (an index of the difficulty of role-switching), presumably correlated negatively with a;

c. The child's success-failure expectancies, both academic and social, as reflected in the child's attitudes and behaviors (SDC, 1972);

d. The positive influence of Head Start, an indirect indicator of the parent's positive attitude toward the child's present school progress (Wolff and Stein, 1967).

3. Estimate of the child's academic potential, relative to his classmates: To the extent that response-biasing is not the sole determinant of replies, this item should reveal the parent's own expectations for the child's school success. Examined from this viewpoint, scores indicating realistic optimism (i.e., at or moderately above the child's actual achievement level) would suggest a good role adjustment as reflected in parent attitudes; avoiding the often cited extremes (Gerard and Miller, 1971; Coleman, 1966) of unrealistically high "aspiration" level and unrealistically low actual achievement expectancies.

4. Estimate of motivation: As with teacher estimates, this judgment attempts to investigate the degree to which the child is involved in the student role, as the parent perceives it. According to Kagan (1971), it is vital for the lower SES child to feel that academic skills are appropriate to his identity and to want to attain them.

Parent Involvement. It is difficult to construe parent involvement in the school setting directly as a child outcome or as a Head Start influence, since so much depends upon the nature of the school system and its efforts to encourage that involvement. It does, however, bear on the question of whether the new role of student is becoming integrated into the child's existing role system. While direction of causality cannot be determined, it is hypothesized that parent involvement is associated with positive attitudes toward the school and its influence on the child's future. *Parent involvement data would be* *collected archivally*, and could include: rate of PTA attendance; volunteering to be "room parent," aide, chauffeur, or chaperone on field trips; positive responses to open-house or parent-teacher conference invitations; and whatever other records of involvement are pertinent for a given school system. Since this information will vary from school system to school system, it would be well to have all the data eventually coded with respect to extent of utilization of involvement opportunities.

Interpretation of Evaluative Constructs. The problem previously raised with respect to teachers' use of evaluative terms arises here in relation to parent evaluations. It is important to determine the extent to which the same characterizations have positive or negative connotations for parents and teachers. Exploration of this question in relation to parents' response structures should not involve use of the large item-pool from which the Classroom Behavior Inventory comes, since the requisite factor analyses could not be performed. However, the question of *similarity of evaluative constructs could reasonably be investigated by means of the semantic differential technique, discussed above*, particularly with the use of audio or pictorial stimuli rather than descriptive terms to be rated. Factor analyses would indicate the extent to which parents' connotative responses replicated those of teachers; of particular importance would be the items found to load on the evaluative factor when stimuli represent minority children. As noted before, these responses are not part of the assessment data but rather help interpret other evaluative data. Thus their collection has low priority.

Observers' Evaluations Based on Standardized Constructs. The nonprofessionals who conduct structured observations of children's action systems will have had considerable opportunity to become familiar with the subjects' behavior styles in the school setting. The observation code categories will in fact have directed their attention to numerous concrete instances of the way each child copes with persons and tasks in that environment. Thus, *at the end of the observation sessions these observers should complete the standardized Classroom Behavior Inventory*. Whether a rating or Q-sort method is chosen for scaling will depend on how many subjects each observer has been assigned. Presumably rating methods of the sort used in the parent interview will also have to be used here, since it is doubtful that individual observers will have become familiar with enough subjects during the observation procedures to use the sorting technique.

The large item pools from which Classroom Behavior Inventory scales are drawn were intended for use as observation checklists (Emmerich, 1971, 1973; Baumrind, 1971, 1972, 1973) by outside observers familiarized with the item content in advance. The present procedure differs only in the following respects. While observers are not asked to attend precisely to the 72-to-200 items on the long checklist form, they are instructed to attend to a variety of target behaviors relevant to the action systems of school children. It is likely, then, that they are in a position to make reliable ratings, although a reliability check ought to be made. Second, only ratings on the shorter inventory will be obtained. Actual event sampling is considered more useful here than obtaining a long list of ratings based on those events. However, the short-form ratings will provide a second source of data concerning the child's enactment of a set of role behaviors previously determined to be important aspects of students' social competence. Comparisons will be made between Head Start and control children based on observers' ratings. It will also be useful to determine the degree to which the teacher and outside observer have similar perceptions of the child's school role behavior. Eventually construction of composite variable values based on both scores might be found to constitute the most representative indices.

In summary, the following sorts of evaluative responses are recommended for collection from parents and other adults who interact with the child in his student role. Parents receive chief attention, being asked first to rate the child within the framework of evaluative dimensions they have provided themselves. Next, they are requested to give summary performance estimates like those obtained from teachers. These measures are to be administered on exactly the same schedule as the corresponding teacher-measures. That is, all judgments are to be collected in the fall for the entire sample, with the first instrument also used for repeated measures on a subsample; summary estimates, however, are needed only once. The parent-generated rating instrument is additionally suggested for use as a pre-measure with a subsample of Head Start and control families during the Head Start year, while exploratory work is proposed to uncover the meaning of evaluative terms for parents. Finally, parent involvement in the school setting (construed as a response to the child's public school role) will be assessed from archival information in school records. Other evaluative responses to the child are to be elicited from observer(s) who record and score his behavior in the school setting. It should be remembered that parents' and outside observers' responses are given least weight in an evaluation of social competency. However, from these sources, along with the data collected from teachers and peers, a comprehensive picture is available of the subjects' role behaviors toward significant others in the proximal social environment.

CHARACTERISTICS OF ACTION SYSTEMS: ACADEMIC

The outcomes and measures of role enactments discussed above deal with the nature of social interactions (viewed both as processes and as the evaluative responses to them) among children acquiring the role of student and significant others in reciprocal roles. Those presented in this section deal with operationally definable characteristics of such action systems, aspects of role performance not specifically interactional but nevertheless thought to be important to the child's achievement of a desired status in the broader social system of which he is becoming a member. The outcomes related to behavioral characteristics are less easily measured and more inferentially related to social competence. They include styles for coping with school-like tasks that are empirically or theoretically associated with academic success. These outcomes are organized in order of priority; those most closely definable in terms of overt action systems are presented first.

Child-task interaction styles are assumed to reflect the child's "everyday effectiveness in dealing with the environment" (OCD, 1973), where that environment is a secondary institution and where effectiveness concerns ability to deal with the sorts of tasks typically encountered in that setting. So construed, the outcomes treated in this subsection partly fulfill the need cited by Walker (1973) to look more closely into "coping styles"--or what the child "does, actively or passively, to handle, organize, accept or influence environmental or internal forces"--as observed in task and test situations or as inferred from classroom reports or records.

Learning styles of children are a second set of academic behavior characteristics. The phrase "learning styles" here means a set of non-cognitive characteristics of children's behavior thought to have an important bearing on achievement in the public school setting. Learning styles influence the way learning opportunities are approached and utilized. Needless to say, not all noncognitive performance styles associated with learning can be independently assessed; in fact, there is not a sufficiently well-established model of the relationship between cognitive and socioemotional factors influencing learning to permit an exhaustive enumeration of the latter. The outcome areas related to learning styles that are discussed here are selected because a general review of relevant literature suggests important relationships between these academic styles and achievement in the public school setting. Furthermore, these outcome areas do not seem to be adequately represented by either teacher ratings or naturalistic event-sampling. Consequently, their assessment requires setting up standardized performance situations in which each child's response style can be observed and measured. The measures would very likely have to be those "experimental" techniques that Walker et al. (1973) note need more refinement before they can be used in a large scale evaluation. As Walker et al. state, it is pointless to try refining paper-and-pencil measures for children this age, so there seems to be no better means than standardized performance situations for getting at the desired outcomes. The measures recommended in this subsection, then, will require substantial pilot investigation, after which their applicability for the entire sample or for a subset only can be determined.

Child-Task Interaction Styles

Of the two behavioral areas described above, child-task interaction styles are the more overt and therefore the more easily definable.

We focus on the child's executive skills, test-taking behaviors, and institutional indices of his successes and failures.

Executive Skills. The term "executive skill" is borrowed from Bronson (1973) and refers to aspects of social competence that are not primarily interpersonal but are task related. These skills are presumed to develop as children practice a variety of strategies in coming to terms with individual learning projects. They represent child-task interaction processes to which institutional reports and records are evaluative responses (in much the same way that interpersonal judgments are seen above as evaluative outcomes of interpersonal interaction processes). Executive skills, then, comprise the aspects of social competence most closely related to cognitive and metacognitive performance. Following Bronson (1973), executive skill

> is defined as skill in choosing and coping with tasks. It requires the ability to select tasks appropriate to one's level of skill, to organize task-relevant materials, to use effective coping strategies, to resist distraction, to notice errors and to correct them or to effectively summon help, to try repeatedly [persist] when necessary, and, ultimately, the ability to reach a chosen goal successfully.

While all these action system characteristics have clear content validity in relation to achieving a desired status within the school setting, they also appear to be the concrete foundations upon which such attitudes as internal control of the learning environment, success expectancy, and academic self-concept are built. Consequently, these overt behaviors are regarded as having far-reaching theoretical significance as well as intrinsic importance.

That task-relevant behavior styles are independently important aspects of social competence, beyond effective interpersonal behavior styles, has already been pointed out. In our discussion of higher-order analyses of behavior rating instruments, we indicated that researchers consistently find a second-order factor interpretable as a task-versus-person orientation, each pole related to different clusters of judgments. Clearly they are independent in that sets of behaviors appropriate and effective in situations that are primarily social will not be appropriate and effective in situations where individual task-accomplishment is the primary goal. It is important to be able to assess task-facilitative behavior strategies, seeing which are most closely associated with objective measures of task success. But it should not be thought that task-orientation and person-orientation are contrary traits, and that a child cannot exhibit behaviors manifesting both. Rather, the orientations come into conflict only in specific situations, so that social competence entails learning both sorts of behavior strategies and also learning to distinguish the situations in which each is relevant. Thus, *there is good reason to recommend structured classroom observations directed specifically at child-task interactions.*

Standard procedures are to be followed in generating a child-task observation system. The same nonprofessional observers used in previously recommended observation studies should be employed for the present assessment, with executive skill data collected during the same set of days but in task-oriented context. Such contexts ought to be representative of the conditions under which most individual learning projects are undertaken; most important, more than one kind of task should be observed per classroom to insure that data are not specific to a particular task. *For these observations, it is recommended that structured individual work periods be chosen in which the children have been assigned a task to work on and complete by themselves with occasional use of the teacher as resource; completion of a work-book assignment is an example.*

More pilot work needs to be done on a child-task interaction scoring instrument than is required for the social-interaction observation instruments. *It is recommended that Bronson's executive skill category system be taken as the basis for such pilot work.* Bronson's executive skill categories include target behaviors thought to be positively or negatively related to task competence. These categories, along with affect categories, need to be extracted from the larger coding system, which includes social interaction coding as well. That they are indeed extractable is implied in Bronson's description (Bronson, 1973, pp. 3-4, 13). The executive skill and related affect categories then need to be supplied with an event-scoring procedure, which should result in a simplification of Bronson's time-sampling method while making it consistent with the scoring techniques recommended for the previous observation instruments.

In its present form, Bronson's observation system is reliable and is recommended by Walker as a potentially powerful noncognitive outcome instrument. If Head Start is effective in providing children with a set of strategies for handling tasks where the notions of self-guidance and individual mastery are involved, then executive skill variables should reflect strong between-group differences. While many social skills might be fostered in neighborhood play situations, it seems doubtful that executive skills important for task competence could be learned outside a preschool setting. *As with the social observation data, it is recommended that child-task interaction data be collected initially near the end of the first month of school for the entire sample. It would also be desirable to make repeated observations of executive skill development for a subsample.*

Test-Taking Behavior. Executive skill in coping with everyday school-like tasks is clearly an important outcome, though it is a difficult one to test. Test-taking might itself be regarded as an example of a typical school-like task, but we have singled it out for special attention both in Chapter 4 and here for several reasons. First, test anxiety and failure threat have been empirically distinguished from general school anxiety and social threat (Sarason et al., 1960; Gerard and Miller, 1971; Katz, 1964, 1968), and both have been seen to characterize lower-status children significantly more often than higher-status children. Moreover, it is generally accepted that while mild arousal facilitates test performance, more extreme concern impairs it and contributes to differential test performance by different status groups. Second, Rotter (1960) emphasizes that the test situation is itself an environmental press affecting the child's test performance. The severity of the press is related to the child's internal needs and to his expectations based on past performance regarding rewarding or punishing outcomes. Thus, the test-taking experience itself contributes importantly to test success as well as to the development of success expectancies and achievement motivations. Finally, the child's test-taking skill is of exceptional importance in that institutions rely heavily on testing for all sorts of evaluation. Walker et al. (1973) point out that all tests measure test-taking ability and motivation to some extent, and do so most noticeably in young children. A compelling corroboration of this point comes from Shipman's (1973) factorial analysis of Head Start longitudinal data collected from 50 tests administered during two years. Two stable factors emerge; the first factor, accounting for 20 percent of all test variance, seems to be test-taking ability. No other single factor accounts for so much variation in dependent measures. Thus, the importance of factors affecting test performance can hardly be underestimated.

We underscore our recommendation in Chapter 4 for the entire sample that, when the first cognitive tests are given, test-taking performance be assessed at the same time using CIRCUS No. 16. This recommendation points up the strong link between cognitive and affective factors in the performance of school tasks. Head Start is expected at minimum to influence the affective component in test variance. That is, Head Start children are expected to adapt more readily to test taking than control subjects, and it is further hypothesized that where such differences occur they will be correlated with differences in actual test outcomes. Evidence for this view is provided in two preschool field evaluations (SDC, 1972; Jensen and Kohlberg, 1966) and in the impressive experimental study by Zigler and Butterfield (1968). The latter study showed significant gains in IQ scores, which were causally related to the reduction of debilitating motivational influences in the test situation as a result of preschool training. They conclude that preschool experience allows children to function better in standardized test situations and thus importantly affects school success potential.

Institutional Indices of Success and Failure. Child-task interaction style in general and test-performance style in particular are seen as two very important processes reflecting the way children are acquiring instrumental role behaviors necessary for effective coping with the school system and eventually with other secondary institutions. These process measures need to be supplemented by evaluative indices provided by the institution reflecting the extent to which, in terms of school records and reports, the child is seen as succeeding in his enactment of the student role. According to Shipman (1973), the large-scale factor analysis of post-Head Start measures revealed a new factor emerging, best interpretable as representing "compliance with social role expectations." Because the factor structure had previously been (and continued to be) stable over repeated test administrations, it was reasonable to regard compliance with institutional norms and rules as a newly emerging and important dimension of social development as a child made the transition into public school. This dimension will be investigated through archival data, scales of early adjustment, and measures of social effects.

Archival Data. Archival data include all records and reports routinely kept on the children in a school system. Such data are referred to as "unobtrusive measures" because they rely on procedures that are in effect independent of the proposed study and require the researcher only to examine existing reports (Walker, 1973). There will be some discrepancy in record-keeping between school systems, but for every comparison of Head Start and control children the data should be equivalent. Moreover, aggregation may be possible through a coding system for giving comparable scores to child outcomes under different record-keeping methods. It is recommended that the following kinds of information be sought:

1. Placement, tracking, or "special class" assignment.
2. Attendance and lateness rates.
3. Referrals to to the school nurse.
4. School success and failure indices, such as nonacademic grades, special awards and demerits, or other evidence of positive and negative role adjustment available in school files.

Information of these sorts will provide mediators of teacher and child academic expectation (item 1), child responses to the academic environment (items 2 and 3), and official responses to the child's behavior in that environment (item 4).

Archival data should be collected only once at the end of the first public school year. However, so that changes throughout the year may be observed for some kinds of data (e.g., absences, nurse visits), it is recommended that information be segregated by time of occurrence, dividing the school year into quarters. First quarter data may be of special interest in distinguishing Head Start from control children, while a repeated measures analysis might indicate when (if at all) such differences disappear.

Scales of Early Adjustment. Many previous researchers have examined the ease with which children adjust to the school role, and all who have done so have found that Head Start favorably affects school readiness. Such early gains, according to Walker et al. (1973), give the child an edge over his classmates, which influences the relationship he establishes with the teacher and their mutual expectations about his future performance. Thus early adjustment may have long-run implications, even though the measures themselves can be expected to yield between-group differences only during the first week or two of school. Because of their short-term applicability, development of these scales is given low priority.

Both Wolff and Stein (1967) and Jensen and Kohlberg (1966) have simply asked teachers for global estimates of school readiness and obtained significant differences for preschoolers, but such a procedure does not identify the kinds of behavior that constitute school adjustment. Stronger differences were obtained by Stearns (1971) and Wolff and Stein (1967) using specific rating dimensions. From Stearns (1971), the following dimensions could be expected to show favorable effects for Head Start populations: cooperation with an adult, aggressive behavior, following directions, ability to pay attention, social adjustment, school attitude, manners. In Wolff and Stein (1967), the following dimensions showed strongest between-group effects: initial adjustment to classroom routines, length of time for full adjustment to classroom routines, behavior toward teacher, behavior toward peers, speech habits, listening habits, work habits.

On the basis of the research reviewed, it is recommended that a specific set of scales based on those used by Stearns (1971) and Wolff and Stein (1967) be adapted for use in the present study. No global estimate need be obtained, since a summary score can be derived from scale ratings. These ratings should be collected during the second week of school (Stearns, 1971; Wolff and Stein, 1967). The rating task should take very little time, and can usefully be compared with the summary estimate of school adjustment obtained later. Given previous research results, it is expected that the more specific items focus on concrete aspects of early adjustment to school routines (e.g., attending to bell, queueing behavior), the stronger the between-group effect will be.

Measure of Social Effects. A nonreactive measure of social salience can be used to index the child's general effect on the classroom environment. The data would be collected from teachers a year after the major evaluation year--i.e., after target children have advanced to the next grade. If any second year measures are to be taken on the entire sample, then this measure should be administered at that time; otherwise, it should be done by mail. What is required is simply to ask each teacher to name the three best and three worst children in the previous year's class. After these names are obtained, the teacher can be provided with the photograph composite used in the children's sociometric task to see how many of the names she can remember. Together these procedures yield a simple representation of the salience of the previous year's individuals, along with an evaluative direction for the most outstanding children. It would be of interest to know whether teachers tended more often to remember Head Start than non-Head Start children; it would be of even greater interest to see whether there was a significant difference in the frequency with which Head Start and control children populated the best or worst student categories. This measure should take very little time and, because the picture composite would already be prepared, it should be inexpensive to give and to score.

In summary, it is recommended that the characteristic behavior of sample subjects toward their academic tasks be assessed in the following ways. First, a structured observation instrument is to be focused at child-task interaction style during individual learning projects. To gauge test-taking style in particular, CIRCUS No. 16, an instrument devised for that purpose, will be administered in conjunction with the cognitive battery. Second, a set of institutional indices of success and failure will be examined. These indices include scales of early adjustment, administered during the second week of the school year; archival data, collected after the year's end; and a social effects measure, obtained a full year after public school entry when the subject has advanced to the next grade.

Learning Styles

The second set of action system characteristics involves styles of learning. Unlike the child-task interaction styles, learning styles are generally covert processes and therefore more difficult to evaluate. We are focusing our attention on the following four significant areas: direction following and task completion, goal-setting and self-evaluation, intentional-incidental learning cues, and curiosity, presented in order of priority.

Learning style measures, which are developed either for the entire sample or for subsamples, are regarded as relevant for only the first post-Head Start year. The actual time of administration is not crucial, although these assessments should be preceded by a few months of school experience. They also need be administered only once. Because early administration is important in the case of many other measures, learning style outcomes could be left for assessment in the second half of the first post-Head Start school year.

Direction-Following and Task Completion. It has already been pointed out that test-taking ability emerges consistently as the first factor in every analysis of Head Start longitudinal data (Shipman, 1973; Walker et al., 1973). Shipman (1973) thinks that this factor primarily represents ability to understand and follow directions, an ability that she regards as underlying successful school performance. Concurring, Baumrind[1] takes it as axiomatic that capacity for work, task involvement, and self-sufficiency are basic to social competence in school. Further, Anderson and Messick (1973) list "control of attention" as a component of social competence; Spivak and Shure (1974) find that such control, which they regard as an aspect of task mastery, is improved by preschool Get Set experience. Spivak and Shure (1974) report that Get Set children are more able than control children to complete tasks alone and overcome obstacles without adult assistance. Such an outcome is related to the Head Start goal of self-discipline and confidence in the learning situation, in implying that the child guides himself through a task without needing continuous outside reassurance (OCD, 1973).

---

[1] Personal communication.

There is good reason to regard direction-following and task-completion, then, as important aspects of learning style, which Head Start could influence. Difficulties arise in deciding how to measure them, however. Teacher ratings are assumed not to be adequate assessments since such ratings become so heavily loaded on general evaluative factors (Williams, 1970a). Some measurement is needed that is more closely tied to specific behaviors. After a careful literature review, two approaches were determined to be most feasible.

The first approach is suggested by the Block and Block test battery (1972), which includes two experimental tasks that seem to involve complex attention-control and direction-following. Both require individual administration. The "competing set" task is most brief, requiring simply that the child repeat exactly what the examiner tells the child to say. After trial repetitions, test sentences invoke competing response sets, which the child must overcome in order to repeat correctly (e.g., the examiner requests "Say 'Do you want to see TV?'" and the child must refrain from answering the question in order to repeat it). The "dual focus" task asks that the child pay attention to story content so that he will be able to answer questions about it correctly and also listen for "clicks" at the sound of which he is to raise his hand. Scoring involves number of clicks and number of content questions responded to correctly.

Although results from these studies are not now available, it seems likely that the competing set task would depend to a considerable degree on verbal ability (particularly, understanding deictic reference in oblique discourse) and would discriminate against speakers of nonstandard dialects (Bikson, 1974b). However, the dual-focus task seems less susceptible to these shortcomings. It will, of course, reflect ability to follow general story content, but that demand is fairly typical of school tasks. Attending to clicks is not, *per se*, representative of any particular school-like demand, but it is the case that children often must attend or try to attend to more than one thing at a time. Here a possibly competing set is involved, but the child needs to integrate the tasks rather than ignore one response set. *Exploratory work is recommended for this second measure, to determine the extent to which it can be regarded as a measure of direction-following.*

The second approach is suggested by Bronson's research (1973), using a structured individual task during which observations of children's work habits are collected. A mastery task could be devised for the present project that would be group-administered, have a goal reachable after following several steps, and involve availability of self-directive cues (e.g., a pictorial representation of instructions to which children may refer if they forget what to do next, and "clues" or partial answer sheets for self-correcting feedback along the way). The task should be scorable for successful completion after administration. It is recommended that the task devised by David Wood and Jerome Bruner (cited in Bronson, 1973) be examined for this purpose. A task specially devised with these features in mind would be the best instrument for detecting the extent to which a child could and would guide himself through a task without continuous supervision.

It is recommended, then, that pilot investigation be carried out with respect to two tasks, the Block and Block dual-focus task and a structured learning-mastery task like that cited in Bronson. Each of these tasks needs to be studied from the standpoint of validity in relation to the notion of direction-following and task-completion in the school setting, as well as with respect to ease of administration. Because of the question of the external validity of findings, a group-administered task of an ordinary school-like nature is preferable. The simplest way of including this outcome measure in the whole sample battery seems to be to require performance of a standardized, specially devised task (scorable for success in direction-following, self-grading and completion) during an ordinary class period. If scheduling complexities rule out this option, then either the same sort of mastery task or the dual-focus task should be administered to sample children separately. Choice of operationalization would then depend on whether it seemed more efficient (on the basis of pilot investigation of both alternatives) to construct a group-administered mastery task or simply adapt the individually administered dual-focus task. *Whichever measure is finally selected should be given to the entire subject sample during the second half of the first public school year.* Success in following directions and completing individual learning projects without continuous outside support is expected to differentiate Head Start from control children.

Goal-Setting and Self-Evaluation. The previous section discussed children's behavior in relation to a task set by the teacher. The present section is directed toward the way in which children set learning goals for themselves and how they appraise such goal-related behavior. According to Kagan (1971), the motivation to master school-like tasks is extremely important to the lower-status child's school success, a view underscored by Butler (1970), Stearns (1971), Baumrind (1973), Zigler (1973b), Anderson and Messick (1973), and Shipman (1973). Because "achievement motivation" and "need achievement" are constructs sometimes treated as cross-situational traits and often without any clear behavioral reference, the notion of self-appraisal in relation to setting and attaining goals is used here instead. However, the definition of "achievement behavior" proposed by Crandall et al. (1962) is worth summarizing. Achievement behavior is behavior directed to attain (avoid) the approval (disapproval) related to competent (incompetent) performance in situations where standards of excellence are applicable. So regarded, achievement behavior is related to the value children attach to intellectual competence, as well as to success expectancy and to self-evaluation standards. Thus, achievement behavior is an outcome centrally involved in a group of constructs, all of which have been regarded both as important for school success and as differentiating higher- from lower-status children.

From the standpoint of achievement behavior so defined, the notions of goal-setting and related self-evaluation processes are important ones to include in the assessment of action system characteristics related to learning. Again, teacher ratings of achievement behavior are not regarded as sufficiently representative of children's behavior, tending instead to reflect status stereotyping (Bikson, 1974a). Consequently, a standardized situation must be contrived to elicit and measure the behavior in question. As before, relevant research literature was critically reviewed. On the basis of that review, it seems clear that some nonprojective behavioral task is required to test goal-setting and self-evaluation and that the task should be related to school learning; further, it should provide some sort of external standard against which the child chooses a goal (unlike the classic McClelland-style achievement tasks) and some way of investigating the child's own standards and self-evaluation practices. Given these conclusions, the best outcome operationalization appears to be an adaptation of achievement behavior tasks used by Crandall et al. (1962) and Weiner (1972) for the present evaluation research, in a way that can be administered to the entire sample. The following steps should be included in the assessment.

1. As with McClelland-style studies, a graded performance situation must be devised. But the performance must be seen as involving intellectual ability. For this reason, a graded series of puzzles is used by Crandall et al. (1962), Weiner (1972), and Block and Block (1973). The use of Porteus mazes should also be investigated, since they might be more efficient to administer although less familiar to preschoolers.
    a. The child will be told he is about to do a puzzle (maze) task, and that there are three kinds of puzzles: those very easy for children his age, those very hard, and some in the middle (Crandall et al., 1962; Block and Block, 1973).
    b. The child will be scored based on the level of difficulty selected (he can be shown examples or just given descriptions, whichever seems to work best). The score represents success expectancy relative to externally provided age norms.
2. After making the category selection, the child is actually presented with the graded series of tasks that allegedly fall within the category he has chosen (in fact, all children will receive the same series). The tasks must range from obviously easy to do to very hard (Crandall et al., 1962; Weiner, 1972) or unsolvable.

   a. The child's minimum achievement standard (Crandall et al., 1962) is shown in this manner. After he has looked at the tasks, the child is asked which one is so easy he would be upset if he couldn't do it because "even a baby could do it." The least difficult one the child would be upset over not being able to do is the minimum achievement standard.

   b. The child is then asked which other of the tasks he thinks he would actually be able to complete successfully, and which he could not do. This procedure yields the child's own success expectancy range (Crandall et al., 1962).

3. The child is then asked to do the tasks, in order. He is also provided with a supply of prizes and told that he may reward himself after he is through working each puzzle with whatever he thinks he deserves (Weiner, 1972).

   a. Level of reward in relation to difficulty level (as perceived by the child and also objectively) will reflect the extent to which self-evaluation is based on internal and external standards of difficulty.

   b. Level of reward in relation to persistence at the task and task outcome at more difficult levels will reflect the extent to which self-reward is based on effort or attainment.

   c. According to Weiner, self-reward behavior can be used to infer causal attribution to oneself of goal attainment. If so, this third aspect of the assessment may be regarded as a basis for inferring perceived internal control of academic outcomes.

4. The possibility of including an attribution test at the end of this series, patterned after the one used by Weiner (1972) with four-year-olds, should be investigated. Pointing to a medium-difficulty puzzle or maze, the examiner should describe the performance of the preceding (imaginary) subject or subjects in terms of ability, effort, and outcome. The child is asked how he, had he been the examiner, would have rewarded the subject since the subject himself was uncertain what his reward should be (the reward range is fixed by the examiner).

   a. Attribution tests typically vary effort, ability, and outcome to determine the relation of reward to causal inference; e.g., reward level can be compared for imaginary subjects who are similar in ability and effort and differ in outcome.

b. While an attribution test of this sort may fall heir to the same sort of difficulties that beset the Gumpgookies test, it may be possible to avoid them because the questions are focused on a concrete task that the child himself has already performed (rather than a series of different imaginary events). Second, the questions are not ostensibly about the child himself, and thus may avoid some of the social desirability bias of the Gumpgookie test. Third, it is expected to be much more brief, aiming only at the child's self-reward behavior under conditions of effort-versus-luck attributions.

It is recommended, then, that pilot work be done to develop a single complex task of the sort described above to access Head Start and control children in relation to level of aspiration, minimum achievement level and success expectancy, and self-reward in relation to performance. The task would presumably have to be individually administered but need not occupy a great deal of time (length of time will presumably be dependent on the actual number of tasks in the graded series). The task could be included as part of a series of tasks that are administered individually; however, it should not be preceded by a task that involves a success or failure component, since an immediately prior performance outcome might be so much in the child's mind as to outweigh all other potential sources of variation in goal-setting and self-evaluation.

Intentional-Incidental Learning Cues and Reinforcement Style. The ability to learn how to learn is one of central importance, as emphasized by Butler (1970), Ziegler and Butterfield (1968), Anderson and Messick (1973), and many others. While very little is known about "learning-to-learn" characteristics, most work in this area has been cognitively oriented and related to information-processing strategies. However, two related phenomena that are socioemotional in nature have undergone investigation and deserve pilot exploration for possible incorporation in the proposed evaluation study. Both action system phenomena are regarded as related to learning style in lower-status children, stemming from a common set toward the teacher's role.

Briefly, lower-status children are seen as much more wary of adult teachers than are higher-status children, a circumstance that makes the learning situation much more stressful for them (Feshbach, 1973; Zigler and Butterfield, 1968). Although causal explanations differ, sources concur in the conclusion that such a set tends to make the lower-status child somewhat more dependent on teacher reinforcement in the learning situation (Ross, 1966; Ziegler and de Labry, 1962; Terrel, Durkin, and Wesley, 1959; Terrel and Kennedy, 1957). This sort of dependency, in turn, has two results detrimental to success in the school situation. First, the child who is so dependent tends not to be selective in the learning process, picking up both relevant and irrelevant cues indiscriminately. Such children may actually learn a greater number of responses, but they will acquire as many responses incidental to the learning task as ones the teacher intentionally instructs (Ross, 1966; Portuges and Feshbach, 1972). The learning of incidental cues, then, actually interferes with attention that could otherwise be given to intentional learning cues. Concomitantly, the lower-status child is seen as needing more reinforcement in the learning situation beyond the sort of reinforcement provided by the information that he has given a correct response. That is, the wary or dependent child is seen as less motivated by correctness for the sake of correctness, requiring material or social reinforcement in the learning situation (Ziegler and de Labry, 1962; Terrel, Durkin, and Wesley, 1959; Terrel and Kennedy, 1957). Ziegler and Butterfield (1968), however, find that nursery school experience allows children to become less wary of adults and more responsive to social than to tangible reinforcers; they also cite numerous investigators who regard the preschool experience as helpful in facilitating a child's transition from dependence on social reinforcers to an interest in being correct just for the sake of being correct.

If Head Start does, indeed, have this effect, it is contributing to socioemotional development in a way very significantly related to potential achievement gains by improving social competence in the typical school-learning situation. Although the two sorts of effects are theoretically related, as the discussion above indicates, they have not been tested together. The same experimental learning procedures cannot easily incorporate both concepts, because intentional or incidental cue-learning is usually tested by not reinforcing either response class and seeing which is in fact learned, while reinforcement efficacy is typically determined by seeing how well a single response is learned given different classes of reinforcers.

Learning of intentional rather than incidental responses has been investigated by researchers interested in imitation and modeling as well as by researchers concerned with influences on school success (Ross, 1966; Portuges and Feshbach, 1972). Operationalization ordinarily involves exposing subjects to a model whose behavior they are explicitly instructed to observe with regard to some future purpose, relative to which some of the model's behaviors are important and others wholly irrelevant. After such a presentation, the subject is given an opportunity to exhibit what he has learned. Subjects' learned behaviors are then scored for number of intentional responses exhibited, number of incidental responses exhibited, and the proportion of total learned responses accounted for by each class.

Although ability to select and learn relevant responses is an important aspect of learning how to learn and is expected to distinguish Head Start from control children, the best way of evaluating this outcome seems to preclude its inclusion in the basic battery. That is, appropriate measurement necessitates individual administration in a standardized performance situation, a task too complex and time-consuming to be used with the entire sample. *It is therefore recommended that pilot efforts be directed at adapting some form of the experimental design described above for assessing the learning of intentional rather than incidental cues in a representative subsample study.* This aspect of learning style could be measured any time during the second half of the public school year.

Children's response learning should also be investigated as a function of the class of reinforcers used by the teacher in an experimental situation. Such learning situations, as represented in recent research literature, typically involve discrimination or concept-switching tasks (e.g., Terrel, Durkin, and Wesley, 1959; Terrel and Kennedy, 1957; Ziegler and de Labry, 1968; Block and Block, 1973). Behavior in these situations is scored in terms of the number of trials required to reach a given task performance criterion. The nature of the reinforcement for correct responses is manipulated, where reinforcers are either material (e.g., candy, prizes), informational (accuracy feedback), or social (approval). Number of trials are analyzed in relation to reinforcement class.

It is expected that Head Start children will be able to learn more efficiently in the accuracy-feedback condition than will control children, holding social reinforcement constant. It would be desirable to have the task group-administered so that it would reflect as closely as possible the circumstances surrounding learning and reinforcement situations in school. In any event, should the predicted difference occur, it would have important implications not only for learning style effectiveness in the school environment but also for conclusions about autonomous achievement motivation. In this way the reinforcement study would provide a valuable supplement to the two previously discussed areas of action system characteristics related to learning.

Pilot study is needed for the nature of the learning task to be used and to select the specific reinforcers whose relative efficacy will be regarded as a learning style indicator. Although such preliminary work is fairly straightforward and unproblematic, resulting procedures may be complex and may not be suitable for group administration. *Thus it is recommended that a test of responsiveness to correctness-feedback versus material reinforcement (controlling for social reinforcement) be adapted from the experimental paradigms cited above for administration to a representative subsample of subjects.* This aspect of learning style could also be assessed any time during the second half of the public school year.

Curiosity. The most complete account of curiosity as a construct involved in all cases of learning from the simplest to the most complex, on both the animal and human level, is found in Berlyne (1960). In human beings, curiosity is thought to be the foundation of exploratory or epistemic behavior stimulated by situations involving a combination of the novel and the familiar and is thus related to the desire to find something out just for its own noninstrumental interest value. From this very brief account, it is easy to see how the construct of curiosity can be theoretically integrated with the other socioemotional influences on learning style discussed above and can be regarded as enhancing the learning process. However, it is not realistic to assume that all or most of what is learned in school is intrinsically interesting (i.e., interesting independent of any instrumental value) to all the children who learn successfully. Conversely, many intrinsically interesting stimuli are not part of school-related tasks, so that curiosity cannot be regarded as a general disposition associated with school success. (For example, lower-status children said to be lacking in curiosity have been found by Anderson and Tindall (1972) to exhibit much more exploratory behavior in relation to their neighborhood environments than do middle-status children of the same age.) Nevertheless, curiosity is treated as an important socioemotional outcome by many researchers (e.g., Block and Block, 1973; Boger and Knight, 1969; Maw and Maw, 1962) and deserves consideration as a learning style.

The best known curiosity measures are patterned after the "curiosity box" used in the Cincinnati Autonomy Test Battery (Banta, 1970). A measure of this sort has been used by Block and Block (1973), who find that it correlates with California Child Q-sort ratings of "is curious, exploring." Such a measure has also been field tested by Boger and Knight (1969) with lower-status minority children; they find it correlates with task initiative but is not associated with task success where success requires impulse control and motor inhibition. It is not clear whether the "curiosity box" as an experimental stimulus elicits *school-relevant curiosity*, or just what the relation between such curiosity and successful academic outcomes is.

A different procedure, and one more school-related, was used by Maw and Maw (1962) to examine the behavior of children high and low in curiosity. Extensive work in a small pilot study as well as a large sample yielded elaborate procedures for identifying children high or low in curiosity. After subjects were so identified, each was presented with sets of stimulus cards. Within each set of three design-bearing cards, one was either unbalanced or unusual (Berlyne, 1960); regarding each set, the subject was asked which he would prefer to discuss, learn about, or hear a story about. (In the pilot, only two design cards were presented at a time, but three-card sets were regarded as better; in the large scale study, "hear a story about" was used as the test question.)

In both the pilot and field study, highly significant between-group differences appeared.

If a curiosity measure of the sort used by Maw and Maw can be developed and validated in pilot study, it is here judged a potentially fruitful measure because of its relation to the sort of epistemic motivation often regarded as relevant to school success. *It is recommended, then, that a measure of epistemic curiosity based on the Maw and Maw (1962) paradigm described above be used for a representative subsample of subjects.* It is expected that, by reducing apprehension in the learning situation and increasing success expectancy, Head Start experience will increase epistemic motivation in treatment subjects relative to control subjects. However, because of the equivocal relationship between such motivation and actual outcomes in school situations, measurement is recommended only on a restricted basis.

In summary, four outcome classes have been recommended for assessment as noncognitive behavior styles related to academic success. Each is to be evaluated in standardized performance situations of the following sorts:

- Direction-following and task-completion are to be assessed in the entire sample by one of two potential methods--i.e., a dual-focus task or a structured mastery task is to be included in the noncognitive battery. Pilot work is needed for instrumentation.
- For measuring goal-setting and self-evaluation, a single complex task must be adapted from a combination of existing research techniques. The measure, to be piloted in preparatory investigation, will investigate level of aspiration, minimum achievement level, success expectancy, and self-reward behavior in the entire sample.
- Influence on learning of cue-relevance and reinforcer class is regarded as an important mediator of academic success. Efficacy of intentional or incidental cues and learning as a function of informational feedback rather than tangible reinforcement will be assessed in two subsample studies.

- Finally, curiosity will be evaluated either using an epistemic-motivation measure or using the curiosity box, depending on conclusions drawn from pilot investigation.

These outcomes classes, ordered to represent increasing learner autonomy in quasi-academic situations, are all expected to differentiate Head Start from control children. Among them, the first two have greatest priority.

CHARACTERISTICS OF ACTION SYSTEMS: SOCIO-INSTITUTIONAL

Outcomes and measures related to social perception and response range have considerable long-term influence in a broad range of future learning and achievement situations, although their assessment poses difficulties for a large-scale evaluation. While these remarks are true of specifically academic behavior styles, they are even more applicable in the social-institutional domain (for which reason this latter domain has lower assessment priority). The notions of "role perception" and "range of response repertoire" are important not only for social competence within the public school context but for success in every sort of social-institutional situation. Problems in their assessment are, however, extensive. First, very little work in role-perception and role-taking research has been done with children, except for sex-role learning (Maccoby, 1959), and those roles are of little interest in the present study. Second, even when adult study is taken into account, most role-related research is exploratory, aimed at finding out what role behaviors typically emerge within a given social context and how they should be described. What experimental work there is usually involves getting the subject to play an assigned role and then seeing how that role performance influences subsequent attitudes or behaviors (Ziller, 1971; Sarbin, 1964). Finally, instruments for large-scale use do not yet exist.

Nevertheless, the outcome area encompassing role perception and response range is of considerable consequence. Sarbin (1964-1968), for example, points out that validity of social role enactment turns first on accuracy of role perception and second on the number of appropriate responses in the respondent's role repertoire. At the same time, valid role enactment is itself necessary for achieving and maintaining a desired social status. Concurring, Weinstein (1969) treats perception of "the role of the other" as the foundation of the development of interpersonal competence. Further, Kohlberg (1969) maintains that persons and institutions, and social concepts generally, are known primarily through role-taking, which he regards as a part of natural cognitive development. Thus, for Kohlberg, the conception of the social self, the social world, and relationships between them--unlike their counterparts in the perceptual world--can be developed *only* through role-taking. In short, the learning and practicing of a number of appropriate role behaviors is clearly relevant to social competence.

The discussion that follows centers on outcomes involving role-taking (divided into spatial perspective, situational perspective, and cultural expectation), and response range (from nonpersonal to interpersonal stimuli).

Role-Taking

Role-taking involves apprehending the perspectives, evaluative sets, and probable reactions to situations that are associated with particular persons by virtue of their holding certain positions in the social ecology. According to Weinstein (1969), such "positional" role-taking is required for children's development as they make the transition from home to school. It should be pointed out that role-taking ability is related to the Piagetian notion of sociocentric versus egocentric perception. Briefly, the latter distinction refers to developmental features of the cognitive organization of the child's experience (Kimbrough and Bikson, 1972). Children at the egocentric stage of cognitive development do not distinguish their own viewpoint from that of others, acting in the firm belief that everyone sees the world exactly as they do. Increasing the range of self-other interactions inevitably provides experiences that conflict with the egocentric structure and require cognitive reorganization that is sociocentric in structure (Piaget, 1948).

Development of role-taking ability, then, is related to sociocentric development and has to do specifically with being able to adopt the position of another in the social ecology. Thus, while development of sociocentric structures is primarily a cognitive phenomenon, it has important links to the development of social competence. It should be emphasized that reference to sociocentric development in cognition generally, and role-taking ability specifically in social comprehension, does not entail the development of any particular set of ethical or social values. Rather, the claim is made that for the child to develop an awareness of the demands, expectations, and values his culture places on various kinds of social behavior undertaken by persons in varied social positions (whatever they may be), he must develop role-taking ability. For convenience, role-taking will be divided into three component outcomes, with potential measures discussed under each.

Spatial Perspective. Actually being able to adopt another person's spatial perspective is perhaps the beginning of the development of role-taking. This ability should be assessed because, to the extent that it constitutes a necessary condition for further social role-taking, it is related to social competence; thus, to the extent that Head Start advances it, Head Start gives its children a foundation for social competence that control children may well lack. Finding an appropriate measure requires only that pilot work be done to single out, among the many available Piagetian egocentrism-sociocentrism tests, the one best suited for children in the age and ethnic categories represented in the projected sample. Two examples are given below.

1. In the Jensen and Kohlberg (1966) study, a simple Piagetian task was used to determine whether the child could distinguish his perspective from the interviewer's. A cardboard house with windows on one surface and doors on the other side was the stimulus. After the whole house had been shown the child, it was positioned between the child and the examiner. The examiner then asked a series of questions (e.g., "Can you see the windows?" "Can I see the door?" and the like) to which only yes-no answers were required. Scores ranged from 1 to 0, "1" indicating all questions were answered correctly.

Jensen and Kohlberg (1966) found that the task was too simple and the range of possible scores too small to permit any differences to emerge.

2. Block and Block (1973) have used a somewhat similar technique with a series of pictures involving characters of varied spatial orientations. Questions involve not only what the child and the examiner can see, but what direction the character is facing and what he can see, what he would have to do in order to see something else, etc.

   While no data are available on the Blocks' sociocentrism test, it seems longer than is necessary and some of the questions about the spatial orientation of the stimulus figure seem more complex verbally than the spatial relation about which information is sought.

It is recommended that some Piagetian egocentrism-sociocentrism task be developed in a pilot study, modeled after those described above and suitable for inclusion in the entire sample battery. Such a task should determine whether Head Start children are developing the necessary foundations for role-taking ability, and whether the Head Start experience has itself provided an interactional basis for those cognitions that control children lack. *It would further be desirable to conduct the egocentrism-sociocentrism test during the Head Start year on a subsample basis.* This test would be relevant for the control population because, while the cognitive structures it involves are important for accurate social perception within the school setting, they do not involve any reference to that setting or to experience-organizations unique to it. Thus, the Piagetian tasks would not be biased in favor of the Head Start subjects, although it is expected that the social experiences afforded by Head Start will result in a significant advantage for that population.

Situational Perspective. In addition to simple spatial perspective, social competence requires the development of sociocentric cognitions of situational roles. According to Sarbin (1968), apprehending situational roles is an organized response of a person to stimuli in a social context, where what is organized is the contemporaneous event plus past experience brought to bear on its interpretation. The perceptual response, then, is the first part of a social act that leads to locating the other's position (based on observation and inference) in the social ecology. Moreover, locating the other's position is complementary to locating one's own reciprocal position in that ecology (Sarbin, 1968; Weinstein, 1969; Cottrell, 1969): We find out who we are and what we are doing through other persons' responses that give social meaning to our acts and define our own role. Thus, learning the situational perspective associated with being a teacher, for example, is an important part of learning to be a student, since appropriate role performance requires learning the expectations related to one's own role and the reciprocal roles closely involved with it. In contrast, failure to locate the positions of the self and others properly will likely result in inappropriate and nonadaptive role enactments (Sarbin, 1968).

Understanding situational perspectives associated with different social roles, then, is an important aspect of social competence. For young children (cf. Weinstein, 1969), the first social system to be dealt with outside the family is the public school. Although many of the outcome classes above (e.g., peer interaction styles, teacher-child interaction styles, child-task interaction styles) examine different kinds of role behaviors regarded as appropriate to the position of student in public school, the present concern is whether the child himself has been able to perceive and differentiate behaviors that are role-appropriate or role-inappropriate with regard to reciprocal role incumbents and tasks in given social settings. This ability may very well be enhanced by Head Start, given Sarbin's thesis that adequate situational role perception involves bringing previous experience to bear on the understanding of a present social stimulus. Besides providing opportunities for role-taking practice, Head Start might provide "anticipatory socialization" (Biddle and Thomas, 1966) by encouraging its members "to adopt the values of the group to which they aspire to belong" (in this instance, the set of successful public school students), which links appropriate and inappropriate role behaviors to group approval and disapproval.

It is therefore suggested that a test of situational role perception be developed suitable for administration during the first post-Head Start year. There is not a large set of instruments from which to choose as in the case of spatial perspective testing, but it is recommended that procedures be modeled after those used by Emmerich (1959) to look at young children's discrimination of parent and child roles. These procedures require a conceptual and an operational step:

1. Emmerich (1959) begins by viewing interpersonal relationships generally as social mini-systems composed of role elements. He examined the family's role composition and found two important bases for role discrimination: Some roles are discriminated by power (child versus adult), and some roles are discriminated by function (mother versus father). The function dimension proved more difficult to articulate, although a major difference emerged between expressive and instrumental functions.
2. The family role incumbents (mother, father, girl-child, boy-child) were translated into stylized figures and presented individually on cards. A set of familiarization procedures was adopted to insure that children understood what the figures represented (e.g., examiner asks, "Who says, 'I'm the mother'?").
    a. Test stimuli consisted of cards bearing all possible paired combinations of stylized figures, randomly ordered to alternate power differences and function differences.
    b. Test sentences were questions based on role relationships (e.g., "Who says, 'You can have it'"; and "Who says, 'No, I won't do it'" etc.), with answers given either verbally or by pointing to the stylized figure ascribed the role in question.

Emmerich (1959) found the results exhibited a developmental trend, with kindergarteners better able to discriminate roles than preschoolers. He also found children had difficulty discriminating the low-power role as a reciprocal role, apparently seeing it only as a lack of power. These results are encouraging in suggesting that role-taking is indeed a developmental phenomenon measurable in Head Start and post-Head Start children and controls. It is interesting to note that within the family social system, the low-power role is difficult for children to identify and define other than negatively. The low-power role in the school social system, it is hoped, would be more readily seen as a reciprocal role involving some specific classes of duties, rights, privileges, permissions, opportunities, and legitimate expectations. It would seem that Emmerich's (1959) procedures could readily be modified in order to get at school role perceptions as well as family role perceptions, allowing these hypotheses to be tested.

A role-taking test involving school figures could appropriately be given only during the post-Head Start year, since control children would have had no institutional experiences before that time. It is expected that control and Head Start children will be of approximately equal ability in role perception relative to the family setting, but that Head Start children will be significantly superior in role discrimination within the school setting and between family and school settings (e.g., when stimulus cards pair parent with teacher). Further, experience with a different set of power and function roles in adults may help the child, by contrast, to perceive more accurately the nature of family roles. In particular, it may supply his perception of his own lower-power position with some defined role-elements. *To obtain a developmental view of role-taking ability as well as to assess the extent to which Head Start influences the perception of family roles, it would be desirable to use an Emmerich-style test restricted to family figures during the Head Start year with a subsample. For this purpose, some additional family-only test sentences may have to be generated during the pilot study (making sure that within-family role relationships adhere to subcultural norms of the subject population). The full family-and-school role recognition test would be administered to the entire sample the following year.*

Cultural Perspective. According to Kohlberg (1969), the term "social" primarily means "the distinctively human structuring of action and thought by role-taking, by the tendency to react to the others as someone like the self and by the tendency to react to the self's behavior in the role of the other." This structuring is guided not only by the constraints derived from specific statuses in social systems such as the family and the school, but also by more generic positions that the society or major culture will assign. These social constraints are often termed "proprietary norms." They cross many situation boundaries and are based on a cultural consensus regarding proper behavior for a person of a given age, sex, and class; such expectations are reflections of salient social values and conformity to them in some degree is probably necessary for acceptable role enactment (Inkeles, 1966; Brewster Smith, 1968).

It was suggested above that in providing "anticipatory socialization," Head Start might be encouraging its members to "adopt the values of the group to which they aspire to belong." Such a statement implies both that cultural value expectations must be met (or at least must not obviously be violated) if the child is to make a successful transition into a majority culture institution, and that these value expectations derive from a culture of which the child is not necessarily a member. Thus, a test of the perception of cultural expectations in a pluralistic society is inherently biased--that is, it must be a test of perception of majority culture expectations, the expectations embodied in the social value perspective of all secondary institutions. Use of a control group, however, allows for exploring the hypothesis that Head Start facilitates majority social value acquisition such that its children are better able to perceive the cultural expectations incumbent on them in the secondary institution than are children who have not had such experience, the difference increasing as a function of the divergence between subcultural and majority-culture values.

Whereas the ability to perceive cultural expectations bearing on role performance is an important aspect of social competence, it is not an easy one to assess. A review of research (Scott, 1969; and Walker, 1973) disclosed a single test of social value acquisition in young children; from that study only the value perception aspect will be detailed here.

1. Scott (1969) conducted a field study of the nursery school as a socializing agency, examining the perception of, compliance

with, and internalization of important social values among lower-SES Australian children as a function of nursery school experience.

a. The first task in the study was to select target values that are explicit goals of Australian nurseries, assuming that by reinforcement, observational learning, and a generally facilitative environment the nursery promotes these values.

b. Values selected by Scott were self-reliance, cooperation, and compliance.

2. After deciding on the key values, Scott elicited from mothers and teachers a set of expectations about how children should behave in situations where these values are relevant. He then devised a set of picture stimuli to illustrate these values in concrete situations, making up a very brief story for each picture.

   a. Picture cards featured an identification figure (a bob-tailed rabbit), with four cards depicting each value area (yielding 12 cards in all).

   b. Questions were devised for each card requiring only a yes or no answer, scored as 1 or 0 depending on congruence with elicited cultural expectations.

Scott's (1969) results showed significant differences in accuracy of perception of cultural expectations favoring children who had had nursery school experiences. He also found significant differences (in the same direction) in compliance with these values and internalization of them, as measured by parent interviews and projective techniques. However, when the latter measures were subjected to analysis of variance using accuracy of role expectations as a covariate control, between-group differences disappeared. These results suggest the overriding importance of accurate perception of cultural expectations regarding one's behavior.

It is suggested that pilot work be done using Scott's (1969) procedures as an example and modifying them for use in the present study. Deciding on the values to be studied and scaling cultural expectations regarding the application of these values in concrete situations will be the most difficult part of pilot investigation. It is not expected that the actual choice of stimulus pictures, administration techniques, and scoring procedures will pose any problems. Consequently, *the test could be administered to the entire sample at any time during the first public school year. In addition, it would be desirable to administer the test to a subsample during the Head Start year, to see how accuracy of perception of cultural expectations develops.* In general, it is expected that the closer parent values are to secondary institution values, the more accurate children's perceptions will be; however, Head Start children are expected to show an advantage over control children, and the advantage is expected to be especially apparent when the discrepancy between parent and institutional values is greatest.

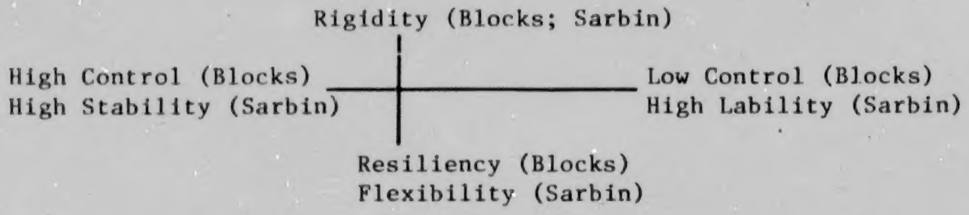
In summary, role-taking is suggested for whole sample investigation in three tasks of increasing complexity: a Piagetian measure of perception of spatial perspective; an Emmerich type of measure of perception of situational perspective; and a Scott type of measure of perception of cultural perspective. If possible, all three of the latter instruments should be administered to a subsample of Head Start and control children during the Head Start year as well, in order to understand increasingly complex role-taking as a developmental phenomenon.

Response Range

The second area of action system characteristics to be examined here is response range. It was noted above that validity of role enactment turns both on accuracy of role perceptions (treated in the preceding sections) and on number of appropriate responses in the respondent's role repertoire. The importance of breadth of response range for psychosocial development is emphatically acknowledged. Sarbin (1968), for example, offers as a "widely accepted postulate" the thesis that the more roles there are in a person's behavior repertory, the better is his social adjustment. Lambert (1963) construes healthy personal adjustment as a function of the amount of freedom an individual has to choose from alternative kinds of behavior. From a cultural perspective, Goslin (1969) points out that socialization itself may be regarded as the process whereby individuals learn to enact various social roles necessary for effective participation in society. Thus, the number of roles in a child's response repertoire is related to the notion of social competence insofar as adequacy of coping with everyday situations is influenced by ability to conceive and deal with those situations in a variety of ways. Moreover, Head Start would be expected to enhance that ability by providing opportunities for learning to enact numerous roles appropriate to positions in the school setting, such as those of student and peer.

While the number of roles enacted by a subject is in principle observable (Sarbin, 1968), there are in fact very few already developed instruments for assessing range of response repertoire, particularly among young children in the school setting. Yet range of conceivable responses is clearly related to ability to solve problems of both an interpersonal and nonpersonal nature and so should be related to both social and cognitive skills. For convenience, discussion of behavior repertoires is arbitrarily divided into two ranges of alternative responses--nonpersonal and interpersonal stimuli.

Response Range to Nonpersonal Stimuli. Generation of alternative responses to nonpersonal stimuli is most fruitfully discussed in relation to the notion of resiliency (Block and Block, 1973) or flexibility in role enactment (Sarbin, 1968). It is interesting that both the Blocks and Sarbin propose the same two-dimensional schema, represented below, for studying the nature of response styles given a variety of task or interpersonal demands:



In both theories, extremes of ego control are regarded as undesirable; but, holding ego control constant, both sources regard resiliency or flexibility as a critical determinant of competence. This construct is interpreted as the ability to negotiate environmental demands without undue anxiety, being resourceful in response to situational stimuli.

At the high end of the continuum, the Blocks say resiliency or flexibility manifests itself as ability to form new accommodations when previously established assimilations prove inadequate. This response style is contrasted with rigidity, a response mode that is stereotypic and unresponsive to unique features of stimulus situations, and is stymied when stimulus structures do not wholly determine appropriate behaviors (Block and Block, 1973). Despite the importance of the construct of resiliency, very little research has been done outside the Blocks' own work to develop workable measures. *Pilot investigation is required to construct or adapt such measures for use in the proposed evaluation*, based on the three outcome subclasses discussed below.

The notion of *boundary elasticity*, or modifiability of conceptual structures, is fairly well worked out theoretically (see Piaget, 1947) and may be measured by means of experimental learning techniques involving concept-switching (cf. Zigler and Butterfield, 1968; Zigler and de Labry, 1962). The discussion of learning styles above makes it clear that learning outcomes might be influenced by other factors beyond resiliency, such as responsiveness to social reinforcement and consequently SES. However, controlling for reinforcement style, any differences between Head Start and Head Start-eligible children in a concept-switching situation might be regarded as a function of differential resiliency. Here it is evident that the assessment of response repertoire characteristics will contribute to understanding response modes in specifically academic contexts as well as broader ones.

As a basis for assessing response resiliency from this viewpoint, the Zigler and de Labry (1962) concept-switching task cited in the reinforcement style discussion provides a suitable paradigm. After preliminary familiarization with sorting procedures in general, Zigler and de Labry (1962) introduce the following sorting problem: The subject is given 25 stimulus cards to sort, the cards displaying five different geometric shapes in five different colors; after the child has successfully performed the sort, he is asked to put the cards together "another way" (i.e., if he sorted on the basis of shape the first time, he would have to sort on the basis of color the second time, and vice versa, although the examiner does not explicitly mention either sorting principle); if the subject fails to switch sorting principles, the examiner points out that he has sorted the cards the same way, and again asks for another way of grouping them. If the subject has failed to switch sorting principles after seven attempts, he is given a score of 8 and the task terminates.

Because techniques for measuring concept switching are so well established, *no difficulty is expected in the adaptation of a performance measure such as that just described for use with the entire sample, preferably to be group administered. Pilot work should also include procedural considerations involved in integrating this resiliency assessment with the subsample evaluation of influences of different reinforcer classes* discussed in relation to learning styles. Head Start children are expected to exhibit greater resiliency on this measure (as well as greater responsiveness to informational feedback).

A second outcome class is suggested by the thesis that resilient (versus rigid) respondents are resourceful in dealing with materials and to respond to unique features of the stimulus situation when those situations are fairly unstructured and leave much to the respondent's own initiative. Such situations might be regarded as *multiple-solution situations* under nonfrustrating conditions. They differ from concept-switching situations in that the latter provide more structure, offering a finite number of alternatives. In the present case, the nature and number of alternatives would have to be provided by the respondent himself. They also differ from barrier situations in that none of the subject's responses are "wrong" or prevented from actualization in the task.

There are at least three kinds of instruments in the Block and Block (1972) test battery that satisfy the conditions presented here: The Sigel unstructured object-sorting task; the parent teaching-strategies tasks; and the "divergent thinking" or "unusual uses" tasks. However, to our knowledge these instruments have not been specifically proposed as resiliency evaluations. Since there is a striking similarity between the theoretical account of resiliency and the kind of behavior involved in these tasks, *pilot research is recommended to establish whether one of the above performance measures is an appropriate test of resiliency* *in an unstructured multiple-solution situation*, such situations being more complex than the kind of situation embodied in the concept-switching task and also more representative of naturalistic problem-solving. *Should a multiple-solution paradigm be found for this purpose, it is recommended for a subsample study during the first public school year.* Of particular interest would be differences between Head Start and control children on the unstructured versus structured (i.e., concept-switching) resiliency tasks, the former regarded as the more difficult and requiring greater flexibility.

Finally, the class of *barrier behaviors*, or generation of responses under conditions of frustration, is included here as a measure of resiliency under the most demanding conditions (Block and Block, 1973). Generation of multiple solutions under such circumstances is different from the generation of multiple correct solutions as described above, because the presence of frustration here adds an emotional condition; the child is in a difficult, thwarting situation and is potentially prevented from attaining a goal. The way a child responds to new thwarting situations will depend upon the repertoire of behaviors he has learned from past successful or partly successful experiences with frustration, as well as from past experiences with the cognitive parameters of the problem. The deployment of a range of responses to cope with the task given the affective condition arising out of frustration is of special interest (Boger and Knight, 1969).

The most popular method of eliciting "barrier behavior" seems to be the puzzle box task used by Block and Block (1972) and Boger and Knight (1969). However, much simpler but highly effective methods of eliciting naturalistic barrier behavior are available in the Blocks' set of unobtrusive measures (1972). Although several barrier situations are provided, the one most amenable for the present purpose is the "stuck drawer" task. Easily incorporated within any individual testing situation, the task requires only that the examiner enlist the subject's aid in getting some pencils (crayons, papers, etc.) out of a drawer while he is preparing the materials for use. The drawer containing the items in question turns out to be "stuck" and the subject's behavior in dealing with this barrier is scored according to details presented in the Blocks' test manual. *Very little preliminary work is required to decide how to standardize the "stuck drawer" situation; it should be included in the basic battery.* The task is expected to differentiate Head Start from control children and to correlate with the previous resiliency measures; of particular interest is the magnitude of those response differences as stimulus situations become more demanding.

Response Range Relative to Interpersonal Stimuli. All that has been said about the importance of resiliency in response to nonpersonal situations holds with even greater emphasis for responses to interpersonal situations. A great deal of theoretical attention has been given to response range in relation to social alternatives. In their classic account of social learning, for example, Bandura and Walters (1963) regard asocial behavior as resulting from the lack of an appropriate response repertoire and suggest that it can be altered simply by providing a range of positive alternatives. Similar conclusions are drawn by Sarbin (1968), who thinks that deviant behavior stems from the absence of opportunities for learning role behaviors appropriate to defined positions. From this viewpoint, the concern evidenced by Bronfenbrenner (1969) and Kohlberg, LaCrosse, and Ricks (1972) for whether children have learned socially responsible behavior patterns characterized by acceptance of reciprocity norms and nonviolent solutions should revolve around the question whether such children have available a range of positive options in their interpersonal problem-solving repertoires.

Although no studies were found comparing nonpersonal problem response-resiliency with interpersonal problem response-resiliency, it seems likely that the two are related (at least the cognitive presuppositions of each would be similar). But introduction of interpersonal elements (like the introduction of frustrating elements in the nonpersonal situation) presumably adds socioemotional complications and, thus, deserves separate study. Further, the explicit emphasis on asocial and antisocial behavior in much of the literature dealing with Head Start as a socializing agent suggests that special attention to positive interpersonal problem-solving alternatives might prove worthwhile. Unfortunately, despite the theoretical emphasis on the importance of a range of alternatives in interpersonal problem-solving, very few measures of this variable exist. Two approaches to this measurement suitable for children in the age range of the present population are suggested here for pilot study; an attempt to develop other measurement approaches should perhaps also be considered.

Sarbin (1968) points out that learning of social roles is associated with *the as-if set* and can be explored by asking a person to behave as if something were occurring; among adults, the exploration often involves posing a set of questions and asking the subject to answer as if he were someone else. This general technique might be adapted to implement Robert Hess's suggestion[1] that a child be presented with photographs of teacher figures in various affective states and asked what he might do if he saw his teacher exhibiting such an expression. While this suggestion is too vague to yield any immediately apparent measurement possibilities, it is feasible to think of using a picture and brief context description as stimulus materials for an as-if task. After stimulus presentation, the subject might be asked to respond to various child behaviors as he thinks the teacher might respond in such a situation. Even more concretely, the child might be asked to respond as his own teacher would to such a hypothetical child-behavior.

In either case, a test of the child's consequential reasoning ability in an as-if situation is provided. Because children's responses will undoubtedly be influenced by their experience, and they have had only one public school teacher, the second course might be best; the same test could then be administered to the child's teacher, with the actual teacher-response used as a criterion for deciding the accuracy of the child's consequential reasoning.

A second test of consequential reasoning, devised by Spivak and Shure (1974), is called the "What Happens Next?" game and has been used in evaluating the preschool Get Set program. The measure is presented as a story-telling game in which the examiner makes up the first part of a very brief story and the child thinks of the ending. There are two series, each involving five such stories. In the first series, one child grabs a toy away from another (only the names of the children and the identification of the toys differs); in the second series, the stories involve a child's having done something without adult permission (again, only names and the specific acts differ). In both cases, the child is asked to think of a different ending for each story. The test measures range of conceivable consequences.

---

[1] Professor Robert Hess, Department of Education, Stanford University.

Stories, story-telling method, probes, and scoring are all quite well worked out by Spivak and Shure. Moreover, the technique proved suitable for a population comparable to the one intended for the present research. However, the test is too long (it takes 20 to 25 minutes) and the need for two sequences is not demonstrated. For the present research, then, *pilot work should investigate the possibility of devising an as-if task based on the suggestion made by Hess*, as described above, for use with the entire sample; perhaps some of the situational measures used by Spivak and Shure would be helpful in this regard. *If, however, development of such a measure is concluded to be infeasible, then one of the two Spivak and Shure story sequences could be administered instead.*

Besides as-if reasoning, a second outcome class is related to as-if reasoning as barrier behavior is related to the generation of multiple solutions, exploring *style of reaction to interpersonally frustrating circumstances*. It is desirable to measure children's responses to interpersonal situations that are specifically problem or conflict situations. Although closely related to the previous outcome, attention in this situation focuses on number of alternative solutions generated to solve an interpersonally thwarting or threatening situation.

For this purpose, Spivak and Shure (1974) have devised the PIPS test, whose procedures are similar to the "What Happens Next?" game. In the PIPS test, there are two uncompleted story sequences, one sequence involving a problematic peer situation (A has a toy and has been playing with it for a long time; B wants to play with it now, but A keeps it) and the other involving a problematic child-adult situation (B broke or otherwise harmed an object valued by his mother and fears his mother will be angry). In both cases, the question is what B can do to deal with his unsatisfactory situation, the score on the PIPS test reflecting the total number of different solutions to the problem sets.

Like the "What Happens Next?" game, the PIPS test takes too long (20 to 25 minutes) to administer. Further, the two parts of the PIPS test are highly intercorrelated, so it is clearly not fruitful to use both. Finally, no evidence of discriminant validity was presented; it is not clear that the "What Happens Next?" game shows anything not presupposed in the PIPS test. *The task for pilot study, then, is to see whether the "What Happens Next?" game empirically discriminates any aspects of interpersonal response style not measured by the PIPS test and to determine which of the PIPS sequences is best included in the basic battery.* Besides seeming more relevant to the subject's range of active coping skills, the PIPS test provides important derived indices such as number of aggressive solutions and proportion of coercive to total solutions. Thus, *the PIPS test is eminently suited for assessing the extent to which Head Start increases prosocial behavioral alternatives and is preferable to the "What Happens Next?" game, should only one of the two measures be included.* Spivak and Shure (1974) found that the Get Set program significantly increased the number of perceived hypothetical consequences in the as-if test, and significantly increased the number of different solutions generated in the PIPS test; it also decreased the force ratio in the PIPS. Similar results are expected to distinguish Head Start from control children in the proposed evaluation.

In summary, social competence with respect to response range was subdivided into two major outcome areas according to the nature of the stimulus involved. Range of response repertory given a nonpersonal stimulus was regarded as a function of resiliency, with response situations organized in terms of increasing complexity. Classes of performance to be investigated included concept switching, multiple correct solutions, and alternatives generated under "barrier" conditions. The concept-switching task was recommended for the entire sample, and a simple unobtrusive barrier situation (the "stuck drawer") was suggested for incorporation with any whole-sample individual test administration. A multiple correct solution task was suggested for subsample study.

The second outcome area, range of response repertory given an interpersonal stimulus situation, was similarly regarded as involving resiliency under circumstances even more complex than those posed by the preceding tasks. Unfortunately the class of appropriate measures was less well developed. Two kinds of performance were singled out for investigation, as-if behaviors and behavioral alternatives given interpersonally conflictual situations. A consequential reasoning task was suggested for development in relation to the former outcome class, while the PIPS test of interpersonal problem solving was recommended for the second. Should the first task not be devisable within the alloted time, another measure--the "What Happens Next?" game--was recommended provided it could be shown to assess aspects of social competency not already indexed by the PIPS test. At least the latter measure, and perhaps an independent evaluation of as-if behavior, are to be included in the basic battery.

ATTITUDINAL CONSTRUCTS

Preceding portions of this chapter have addressed themselves to the assessment of action systems or characteristics of such systems. The final section concerns attitudinal constructs thought to be importantly related to children's effective role performance in the social ecology, with particular attention given to attitudes empirically or theoretically linked to successful participation in the academic environment. It is not without misgivings, however, that the question of attitude measurement is approached at all. First, the relationship between attitudes and behavior is unclear. The consensus in contemporary social psychology, moreover, seems to be that behaviors generally cannot be inferred from attitudes, although the converse direction of inference is promising (Jones and Gerard, 1967; H. H. Kelly, 1967). Thus, the data collected regarding behavior as outlined in the preceding parts of this chapter should provide better indices of related attitudes than would most attempts to measure attitudes apart from such behaviors. Second, in striking contrast to the multitudes of adult attitude instruments, "there is a dearth of instruments suitable for young children" (Walker, 1973). Further, adult instruments cannot be adapted for use with young children, since such instruments are invariably paper-and-pencil measures. To recommend attitude measurement, then, is to recommend extensive initial pilot work whose fruitfulness is uncertain. Consequently, only a very few attitudinal constructs thought to be especially significant in the development of social competence among Head Start children will be treated in this subsection: school attitudes and self attitudes. In terms of the chapter as a whole, measurement of attitudinal constructs has lowest priority.

School Attitudes

School attitudes form the class of attitudes most confidently linked to the successful performance of school roles. According to Stearns (1971), the child's school attitudes are intrinsically important whether or not they can be associated with any specific class of performance variables. Kagan (1971) agrees that, especially for the lower-SES child, intellectual motivation is an important outcome. Removing the issue from the instrumental domain, Kagan (1971) contends it is a cross-cultural expectation that every child should wish to be intellectually competent and should expect to attain that goal; from this viewpoint, intellectual competence is a general societal value, and no child, regardless of status level, should be prevented from acquiring it. Sarbin (1964) looks to performance. He points out that forced compliance with role requirements incompatible with self-conceptions will not produce long-term changes because the self is not involved in the role; but self-role congruence facilitates the kind of motivation that does lead to effective role performance. This viewpoint would regard favorable school attitudes as important mediators of competence in the student role. That such attitudes are instrumentally important for school success is a broadly accepted thesis (e.g., Zigler, 1973) and needs no discussion here.

Despite general acceptance of the proposition that the child's favorable involvement in school roles promotes effective performance of those roles, there are few generally accepted measures of such attitudes. A review of relevant research leads to the suggestions, described below, for the investigation of school attitudes.

First, an individual interview should be conducted attempting to bring out the most important school attitudes, relying on procedures developed by psycholinguists for eliciting spontaneous speech from young respondents (e.g., Cazden, 1967). While extensive individual interviews would not be appropriate for the entire sample, their use with a subsample should be helpful in providing a criterion against which other attitude measures can be interpreted and evaluated.

For the sample as a whole, a more structured verbal self-report is probably the most feasible instrument. Among potential verbal self-report measures reviewed, most are too complicated for the present subject population. However, two are recommended for exploration in pilot research:

1. The Primary Academic Sentiment Scale (PASS, reviewed in Walker, 1973) can be used with subjects as young as four years and has been used to evaluate effects of Title 1 programs on school attitudes of young children (Dowd and West, 1969). The instrument is composed of items read aloud to subjects and can be group administered; items request information on children's preferred activities, attitudes, and behaviors as well as his parents' activities and behaviors. The test is administered in two sessions. While age-quotient norms are available and validity is rated "fair," the reliability of the scale is "poor" (Walker, 1973); internal reliability coefficients are 0.70 and 0.80 for kindergarteners, and lower for preschoolers.

   It is reasonable to think that scale reliability could be improved by making minor revisions in the instrument. Most important, group administration probably detracts from reliability considerably. Further, the test is probably too long and could be advantageously shortened. In particular, items referring to parents' activities should be omitted, parent information being more easily obtainable through other sources. Some items referring to children's behaviors might also be omitted, much behavioral information being directly obtainable through the observation methods previously recommended. Only items related to preferences and attitudes of children should be retained. Its fair validity provides the strongest reason to explore the PASS instrument.

2. Minuchin et al. (1969) have used a sentence-completion scale to measure school attitudes among children who have been involved in differently structured school programs. After children have been familiarized with completion procedures, they are asked to complete the following sentences:

   One good thing about school is __________________________.
   When the teacher leaves the room, _______________________.
   When you are late for school, ___________________________.
   When the teacher asks for quiet, ________________________.
   I try not to ____.

   Sentence completions are rated on a four-point scale (negative, ambivalent, conforming, and positive identification). These scales reflected significant between-group differences interpreted by Minuchin et al. (1969) and by Stearns (1971) as indicating that some school programs lead children to greater student role investment.

   In view of Minuchin's success in detecting with these measures the sort of attitudes the present evaluation hopes to investigate, we recommend pilot investigation of their use with younger children, since the youngest of Minuchin's subjects were third-graders.

Borrowing from psycholinguistic techniques, it is recommended that the "alligator game," so successful in eliciting structured verbal responses from children as young as three years, be explored for use as a medium for collecting the verbal self-report data outlined above (McNeill, 1970). The alligator game uses two very attractive fuzzy green hand puppets, a crocodile and an alligator. For familiarization procedures and instructional purposes, the interviewer initially manipulates both puppets. For psycholinguistic testing, the crocodile utters a sentence and the alligator responds with the desired transform (e.g., crocodile: "The boy chased the dog"; alligator: "the dog was chased by the boy"). After enough examples so the child clearly understands what is going on, the interviewer says, "Now would you like to be the alligator?"; the child is given the alligator puppet, and the crocodile continues supplying stimulus sentences.

The alligator game format could easily be adapted for any sentence-completion test; for this purpose, the crocodile would begin test sentences involving school attitudes and the alligator would finish them. Children seem to enjoy this unobtrusive and nonthreatening procedure. It is recommended that pilot work be directed at selection of the most sensitive items from the PASS or Minuchin et al. school attitudes tests for use with an alligator game response format, to be administered individually to the entire subject sample during the first public school year. To the extent that Head Start succeeds in facilitating the transition to public school for lower-SES children, it is expected to improve their school attitudes in comparison with the attitudes of control subjects. Attitude differences in the predicted direction would signal greater consistency for Head Start children between self-conceptions and the requirements of the role of pupil.

A second measure focuses more directly on self and academic-role congruence. Devised experimentally by Crandall et al. (1962), The Children's Achievement Wishes Test is useful for exploring academic preferences.

This test uses 18 picture pairs as stimuli, pictures depicting children engaged in various academic and nonacademic activities. The child is asked, regarding each pair, "If you could have your wish, which one would you like to do especially well?" The test yields 18 forced-choice responses interpreted by Crandall et al. as representing the value a child attaches to intellectual competence.

Crandall et al. used this test with children as young as first grade from varied status backgrounds. They found achievement wishes scores correlated significantly with scores on a measure of internal control of academic outcomes and also with the amount of free-play time children were observed to spend in intellectual pursuits; these results provide a basis for thinking the measure is a valid one, although the forced-choice format does not allow determination that a child actually likes one of the choices independently of the fact that he prefers it over the other alternatives. If exploratory work establishes that his forced-choice picture test is a workable measure for use with the present research population, it should be included in the major test battery; pilot study should also determine whether the test could be group-administered. This instrument, in whichever form it is administered, is expected to show Head Start children preferring academic settings significantly more often than the control population.

A number of assessments of a more experimental nature are promising for use in focused studies. Among them are a behavioral counterpart of attitude rating based on the classic Byrne-Nelson similarity-liking model (see Gaynor, Lamberth, and McCullers); an adaptation of the World Test (Block and Block, 1972), where attitudes are judged on the basis of themes emerging in fantasy play; and finally the videotaping of children who are engaged in watching school scenes for subsequent coding to yield indices of school attitudes. Evaluations of the sort just mentioned, while shedding light on the nature of attitudes lower-status children develop toward public school and toward the institutional setting generally, cannot be recommended for use with the entire subject sample. They are discussed more fully in Chapter 10.

The approaches to assessing school attitudes enumerated above produce few definite conclusions. *It is recommended that an interview situation be used with a subsample of subjects to yield a criterion against which the remainder of the attitude assessments can be evaluated. Pilot investigation is expected to yield a verbal self-report of school attitude and a forced-choice measure of academic achievement wishes suitable for inclusion in the basic battery of tests for the entire sample; the other measures would probably be applicable only in focused studies. Assessment of school attitudes may be undertaken any time during the first public school year.* It is expected that Head Start experience will serve both to make academic roles seem more appropriately related to the child's own identity and to facilitate entry into the institutional setting, thus creating more favorable school attitudes.

Self Attitudes

Beyond attitudes explicitly concerned with aspects of the school situation, those most frequently treated in relevant research are attitudes related to the self-construct. Assessment of aspects of the self-construct is difficult because there is neither a firm empirical tradition nor a widely accepted theoretical model establishing the parameters of vital and healthy self-constructs in children, particularly in a culturally pluralistic context (Goodchilds, Green, and Bikson, 1974). As Proshansky and Newton (1968) point out:

> [T]he relevant literature contains a confusing assortment of terms which refer to the individual's beliefs and feelings about himself.... While these differences in terminology reflect differences in theory and method, the differences are far from clear-cut. Furthermore, even when theorists or investigators actually employ the same term, they are by no means always in agreement as to its meaning.

Lacking a single validated model of the self-construct in children, the present report examines only a small set of self attitudes that have been used to indicate psychosocial adjustment and have some *prima facie* relevance to social competence among Head Start children in the public school situation.

Self-esteem has been an attitudinal construct of considerable interest in research on school adjustment, particularly among minority and lower-status children. Proshansky and Newton (1968) comment that the importance of a preponderance of favorable judgments covering many dimensions of self is treated as axiomatic in most of the sources they review. A relationship between favorable self-evaluation and school success is established by many researchers (e.g., Lambert, 1963; Hauser, 1971; Sarason et al., 1960), although the two are presumably reciprocally influential. Further, Walker et al. (1973) point out that there is a need to measure self-concept because so many programs incorporate the improvement of self-concept as an aim. This comment applies specifically to the Head Start program. Shipman (1973) sites self-esteem as a variable of major future concern in Head Start evaluation. However, these same sources acknowledge considerable difficulty in measuring self-esteem, particularly among subjects in the proposed age range (Walker, 1973; Walker et al., 1973; Shipman, 1973; Emmerich, 1973). It does not seem advisable to recommend extensive pilot work on constructs that are currently the subject of a great deal of independent research by psychologists. The recommendations below, then, make use only of what has seemed the most promising recent research.

An individual interview should be conducted with a subsample of subjects to investigate the most important self-evaluative dimensions. Pilot investigation is needed to determine both item content and the best method of eliciting responses. Presumably it would be efficient to incorporate these pilot efforts with the pilot research on interview measurement of school attitudes. As with school attitude measurement, self-attitude measurement may derive criterion variables from interview results.

Although the individual interview is capable of differentiating varied aspects of self-concept, it is relevant for only small subsample studies using well-trained interviewers. A measure recommended for inclusion in the basic test battery is the Children's Self-Social Constructs Test based on Ziller's work (Ziller et al., 1969) as adapted for use with preschool children (Walker, 1973) and available through the Educational Testing Service. As Henderson, Long, and Ziller (1965) see it, it is extremely important to have nonverbal measures of self-constructs since verbal ones are so visible, so susceptible to response biasing, and often dependent on age or verbal ability; they assume subjects to be quite articulate about their own self and social constructs. These considerations constrain the usefulness of many such measures among adults as well as among children. The Self-Social Constructs Test, in contrast, uses gummed labels and stick figures to stand for the self and significant others, with circles representing their position in the social environment.

The self-esteem measure is very simple, involving only a page on which there is a vertical column of five circles. The subject is given a gummed label, which he is told represents himself; he is further told that the circles are children and is asked to pick one to be himself and to paste his label in it. The self-esteem score is based on the circle selected, the highest score given for the highest position in the column (here self-esteem is regarded as a value attached to the self in comparison with others). The review of this measure in Walker (1973) suggests it is reasonably valid and reliable. Moreover, the extensive field testing conducted by Boger and Knight (1969) indicates that it is reliable with ethnically diverse Head Start populations. Specifically, they comment that the children, contrary to their initial fears, have no difficulty treating labels as "self" symbols; rather, the task is easy to perform and not too abstract. Concurring, Miller and Dreger (1973) give a positive evaluation of the Ziller test in their extensive review, noting that it is a reliable, unobtrusive, and theoretically well-founded instrument.

It is therefore recommended that at least the Ziller self-esteem measure be included in the test battery for the entire sample. In addition, other items might be chosen for inclusion from the self-social constructs test, although if the entire test is given it might prove too long (Boger and Knight, 1969). *The measure of the child's social distance from his teacher (scored on the basis of how many circles he allows to stand between the teacher-symbol and where he locates himself) is especially recommended* because it has an obvious bearing on school attitudes and because it has been found to yield significant differences between achievers and nonachievers, and between Head Start populations and middle-class controls (see Walker, 1973; Henderson, Long, and Ziller, 1965). It is recommended that pilot work be undertaken to produce a shortened version of the Self-Social Constructs Test including at least the measures of self-esteem and distance from teacher, along with however many additional social construct measures (e.g., social interest, minority identification) are feasible for inclusion in the entire sample test battery.

Multiple Role Integration

The last attitudinal construct proposed for study is so little researched that it does not have a popular name, much less an available instrumentation. Yet it is clearly a self-construct of major importance, referring to the degree of success with which an individual integrates the roles he is required to enact. It has already been suggested that children entering the public school situation are obliged to acquire a new role repertoire related to effective coping with secondary institutions, and that for Head Start-eligible children the transition often involves a change in cultural context as well. The transition thus poses a special challenge for such children, who are faced with the necessity of coming to terms with diverse and potentially conflicting role expectations in order to emerge as socially competent in a culturally pluralistic society. This notion is represented in Hauser's (1971) definition of healthy psychosocial adjustment as the "integration of self-images and social-role-images over time in such a way as to allow for the fullest self development" or at least for an "adaptive self development" of the individual. The emphasis on successful enactment of multiple roles is represented explicitly in Anderson and Messick's (1973) paper and in the Rockefeller University workshop, where concern focuses on social code-switching and on integration of differentiated self-aspects into an identity. A narrower interest in the congruence of academic role behaviors with the self-concept of lower-status children is expressed by Kagan (1971), who regards such congruence as intimately involved with intellectual mastery motivation. Finally, concern over the "alienation" of the lower-status family from the broader community (Shipman, 1973, Zigler, 1973b) implicitly acknowledges that such families find it difficult to integrate relationships with secondary institutions into existing activity patterns.

Role-theoretic literature provides a framework within which to view the situation of the Head Start child acquiring the role of public school student. Accounts of role acquisition and role transfer suggest three ways in which individuals may respond to the necessity of incorporating a new role repertoire (Weinstein, 1969): (1) the new role may be consonant with internalized norms, in which case it will simply be added to existing roles in the behavior repertory with very little difficulty; (2) restructuring of orientations (norms or role definitions) may be required to resolve dissonance between the new role and internalized norms; or (3) defense mechanisms or deviant adjustments may be invoked to handle unresolved dissonance (as alienation, withdrawal, antisocial or amoral behavior). While the measures of role-taking mentioned above focused on aspects of valid role enactment, the question of multiple role integration is often construed negatively in terms of potential role conflicts. Sarbin (1968) points out that indefiniteness and diffuseness of role expectations tends to characterize the condition of persons mobile from class to class or culture to culture. Role conflicts arise when expectations related to the enactment of one role are discrepant either with the enactment of another role in the behavior repertory or with the dictates of the self-construct (role-role conflict and self-role conflict, respectively; Biddle, Twyman, and Rankin, 1966; Biddle and Thomas, 1966). Such conflicts are experienced with anxiety (Sarbin, 1964; Sarason et al., 1960) and, if unresolved, typically lead to behavior problems (Sarbin, 1964; Weinstein, 1969; Biddle, Twyman, and Rankin, 1966).

The model for explaining emergence of behavior pathology stemming from unresolved role conflicts is based on three components: (a) Antecedent events are seen as stressors in the form of role demands the person cannot find a way to satisfy; (b) the intervening process is regarded as a time of arousal accompanied by anxiety, cognitive strain, and possibly physiological perturbation; (c) adaptation takes the form of role enactments designed to validate the occupancy of irregular, autistic, or unconventional statuses, or statuses with minimal obligations (Sarbin, 1964; Hauser, 1971; Dohrenwend and Dohrenwend, 1969). This model is perhaps a workable one for viewing the withdrawn, underachieving, defiant, or antisocial behavior that sometimes characterizes the entry of lower-status children into the public school setting (Hess, 1974; Bronfenbrenner, 1969; Kohlberg, LaCrosse, and Ricks, 1972). Unfortunately no equally well-developed models of *successful* multiple role integration are available, although the notion of interpersonal resiliency discussed above is a start in this direction.

Given the importance of successful multiple role integration for social competence as opposed to social dysfunction in Head Start children, focused study is recommended to explore the following area. First, *it is necessary to develop a description of the number of social positions occupied by members of the subject population.* Positions can be defined in terms of clusters of role behaviors (e.g., family roles versus school roles versus street roles). Sarbin (1968) suggests that the methods developed by sociologists or cultural anthropologists would be appropriate for such an endeavor. In addition to a descriptive account, it is important to include information about behavioral qualities, styles of role performance, preferences, attitudes, and values thought to accompany occupancy of different social positions and the standards of different social groups who judge the competency of the position's occupant. Here potential sources of role conflict should be identifiable. Second, *investigation should pursue the ways in which potentially conflicting roles are handled by children who are successful and unsuccessful at the task of integrating them*; such an undertaking involves identifying potential conflict cases and scaling the adequacy of the response. If Head Start were effective either in helping children avoid pathological solutions to social role conflicts or, better yet, in facilitating resilient, growth-oriented styles of multiple role integration, its contribution to long-term social competence would be inestimable.

Chapter 6

INDEPENDENT VARIABLES

TREATMENT VARIABLES 236
  Amount of Treatment 237
  Treatment Environment 237
  Treatment Events 243

CONTROL VARIABLES 253

BACKGROUND VARIABLES 256
  Child Background Variables 256
  Family Background Variables 258
  Teacher and Teacher Aide Background Variables 259
  Center Background Characteristics 260
  Site (Catchment Area) Variables 261
  Kindergarten And First-Grade Variables 262

Chapter 6

INDEPENDENT VARIABLES

This chapter discusses three categories of independent variables and their measures:

- The *treatment variables* provide a description of those aspects of Head Start programs that are most likely to make a difference in children's outcomes.
- The *control group variables* are those variables that specify the control conditions, and they will help in interpretation of between-group differences.
- The *background variables* are those characteristics of the subject sample (such as age, geographic region, and ethnicity) that are expected to explain some of the effects of Head Start on the outcome variables.

These three categories of variables are presented in order below, along with the reasons for selecting them and the manner in which they are to be measured.

TREATMENT VARIABLES

Three classes of treatment variables are selected for inclusion in the Head Start evaluation study: (1) the *amount* of treatment, (2) the treatment *environment*, and (3) the *events or processes* that make up the treatment. We include treatment variables in the proposed evaluation because they can be useful in determining Head Start program characteristics that are likely to make a difference in outcomes for children. We have formulated recommendations around those independent variables that appear to be linked to the most significant or consistent results across studies. Treatment aspects are the independent variables of greatest interest from the standpoint of evaluation and policy formation in the area of child development, and they are considered below.

Amount of Treatment

The first set of classroom instructional variables specified as independent factors is the *amount of treatment received*. The most compelling reason for including this class of variables is its face validity, with regard to both expected outcomes and policy relevance. That is, it is appropriate to hypothesize that, in whatever outcome areas effects may be sought, the magnitude of gains will reflect the magnitude of the input, other things being equal. Thus, sample children in two-day programs will be expected to show dependent measure scores systematically different from those of children enrolled in classes that meet every day. Similarly, half-day programs and full-day programs are assumed to yield different results.

Several previous Head Start studies have shown significant differences in results related to input magnitude. The Westinghouse study (Cicarelli et al., 1969), for example, confirmed the greater efficacy of a full year program over summer school programs by examining posttest scores on a number of dependent variables. The study by Jensen and Kohlberg (1966) indicates that substantially different effects occur as a function of differences in amount of treatment. Their research showed that summer Head Start sessions were able to produce gains only on socioemotional dimensions related to school adjustment. Longer programs produced both those gains and cognitive gains. Amount of treatment, then, is the first program characteristic selected for study.

Treatment Environment

The second characteristic, *treatment environment*, includes physical and social aspects of the educational setting that are somewhat static. The treatment environment comprises the curriculum model or plan, the teachers' attitudes toward its implementation, and the materials by which it is implemented. These variables specify the context in which Head Start children spend their time and supplement the durational variables described above. Three components of the treatment environment are distinguished below: a curriculum model, teachers' attitudes, and the physical-social setting.

1. *A curriculum model, reflecting varying emphases on Head Start goals of compensatory education and social development, should be specified for each Head Start center studied (or for some proportion thereof).* One independent variable that seems to capture important information about the nature of the educational setting of Head Start centers is the content of the curriculum and its emphasis, as fostered by the center directors or teachers. Weikart (in Cazden, 1972) and Grannis (in Cazden, 1972) have classification schemes that might be used to distinguish preschool curricula. The Weikart scheme distinguishes curricula in terms of different emphases on academic skills and social facilitation:

a. Programmed (e.g., Engelmann-Bereiter).
b. Open framework (e.g., DARCEE, Weikart).
c. Child-centered (e.g., Montessori, Bank Street).
d. Custodial.

Grannis has classified three learning goals in describing an educational setting:

a. Type 1 learning--no individuation; transmission of common knowledge, culture.[1]
b. Type 2 learning--partial individuation; internalization of concepts and skills.
c. Type 3 learning--major individuation; cultivation of individual exploration, application, and expression.

Except for Weikart's custodial category, which appears to offer no (or only incidental) education, it seems likely that different program emphases produce quite different child outcomes. A review of the literature seems to confirm that there are relationships between curriculum characteristics and child gains. Most previous studies have been concerned with the analysis of particular models with different program attributes--for example, those used in the Head Start Planned Variation (HSPV) experiment and Follow-Through project. Since only a small number of sites were involved in both the HSPV and Follow-Through projects, some difficulty may be incurred in finding Head Start program variations that approximate any given scheme. Nevertheless, judging from the natural variation found among Head Start centers and the HSPV models implemented in over a dozen sites, it is likely that "curriculum model" is a reasonable method by which to characterize centers.

---

[1] Individuation connotes individual as opposed to mass activity. An example of Type 1 learning is group singing or recitation.

If, in fact, different Head Start curricula produce different outcomes for children, the federal, regional, and local levels of the program can use this information to select different curricula for different goals for children. Examples of possible curriculum-outcome relationships might be the following. Classes categorized at the skill-training end of the continuum of program emphasis will show certain types of cognitive gains, as well as certain types of social-personal gains. Classes categorized at the child-centered end of the continuum (custodial models being in a separate category) will show quite different types of social-personal gains, as well as different types of cognitive gains. Classes that achieve a balance among *all* three types of learning will produce maximum cognitive and social-personal gains. In all three instances, empirical evidence would have to be able to describe the nature of the cognitive and social gains that would be characteristic of each emphasis. Considerable innovation and study are needed in this area. Earlier thinking has been focused primarily on school learning (Grannis' Type 2) and individual development (Grannis' Type 3), often underestimating the significance of community learning as suggested in Grannis' Type 1 learning.

Before a curriculum plan or model can be used as an independent treatment variable in a large-scale Head Start evaluation, it is necessary to establish reliable methods for categorizing Head Start programs. Two complementary schemes should be considered for classifying curriculum variations. In the Weikart classification, current practice is followed; in the Grannis scheme, field-testing of the reliability of typological judgments is required. The Grannis scheme differentiates *how* something is learned, not *what* (e.g., cognitive versus social skills). In other words, where the Weikart classification of programs is a continuum from cognitively to socially oriented programs, each *type* of Grannis' scheme allows a continuum from cognitive to social learning. It is expected that Weikart's classification will generally correlate well with the

Grannis system--i.e., most programmed and open framework programs will emphasize Grannis Types 1 and 2 learning, while most social development or child-centered programs will stress Type 3 learning.

Procedures for assigning scale values on the dimensions recommended become a concern second only to the choice of classification scheme itself. Cazden (1972) comments that curriculum classifications are "based on the rhetoric of the curriculum designers, not on what actually goes on in their classrooms." Therefore, any measurement of curriculum should reflect classroom processes more closely than it reflects intensions. There appears to be a sizable gap between the design of curriculum objectives and their implementation in the classroom even under controlled conditions. Specific curriculum model variations may be difficult to identify in the variations occurring naturally among Head Start centers. A recommended solution is to obtain two or three inputs: the teacher's own perception of her classroom emphasis in terms of the adopted classification scheme; an observer's perception of that emphasis (using the same observers who score classroom events, as discussed below); and possibly the center director's perception of the curriculum plan or model. Preliminary investigation is needed to determine the best way of combining these three sources of information (and how to weight their judgments) to yield a curriculum variable for differentiating treatments.

2. *Teachers' attitudes toward the implementation of the curriculum should be assessed for each Head Start classroom.* Teachers' perceptions of the role of compensatory education, of their jobs in the center, of specific goals they adopt for their classes, and of the methods used to reach those goals affect how a curriculum model is implemented. Information about these factors is regarded as an important supplement to curricular typology and is potentially capable of differentiating between Head Start centers with similar basic models or plans. Data on teachers' attitudes toward curricula should help fill the gap cited by Cazden (1972) between curriculum design and what actually goes on in the classroom, supplying teacher-related intervening steps.

The perceptions of teachers as Head Start educators can intervene between curriculum design and classroom activities, and those perceptions have been related to children's outcomes in several ways. First, the relationship between teachers' attitudes and expectations and children's actual achievement has been studied by many authors, most notably by Rosenthal and Jacobsen (1968). There has been some research on how teachers' attitudes and expectations are translated into classroom objectives and techniques. These objectives and techniques provide concrete links between teachers' implicit views about their roles as Head Start educators and children's outcomes and thus merit further investigation. Second, the approaches to curriculum implementation by the Head Start teachers may be associated with their backgrounds. For example, disadvantaged children's outcomes are negatively associated with duration of the teacher's paid experience with disadvantaged children (SDC, 1972). Such a relationship is difficult to understand apart from assumptions about teaching attitudes or teaching techniques. It would be important to test these hypotheses from a policy standpoint. No program can alter the extent of previous experience, but it would be possible to change the influence of that experience by providing alternative viewpoints and methods for such teachers. Finally, independent of attitudes and expectations, reinforcement style has been established as an important contributor to children's outcomes (see Lamb, 1965; Chapter 5, above).

Teacher's curriculum implementation will be measured by four subscales of the CIRCUS Educational Environment Questionnaire (CIRCUS No. 17). First, the teacher's general attitude toward preprimary compensatory education and the nature of instructional processes will be measured by items 22 through 63 ("educational viewpoints"). Next, the teacher's feelings about the specific center in which she is currently a preprimary educator are reflected in items 8 through 21 ("job"). Third, specific classroom objectives entertained by the teacher are indexed in items 77 through 96, together with their priority from her point of view ("educational objectives"). Finally, the techniques used by the teacher to change children's behavior are assessed in items 64 through 76 ("techniques"). This self-report questionnaire is to be filled out by the head teacher in each sample classroom and should provide a fairly detailed account of the way the curriculum design is implemented and contribute substantially to knowledge about the treatment environment for evaluation purposes.

3. *The physical-social setting in which the treatment occurs should be assessed for each Head Start classroom.* Although there is not a great deal of research literature linking the physical-social setting to children's outcomes, certain aspects of that setting are considered to be important treatment variables.

First, the *OCD-Head Start Policy Manual* (1973) explicitly acknowledges the necessity for cultural recognition of minority children's backgrounds and has as its goal the enhancement of ethnic identity. The most obvious way in which Head Start centers carry out this policy is by displaying and using culturally relevant materials in the classroom. In addition, special instructional facilities and materials (e.g., a science learning area) have been thought to influence the kind of cognitive gain made among preschool children. Even the use of television for regular "Sesame Street" viewing has been associated with children's outcomes. Ascertaining the availability and use of these elements in the preschool behavior space thus is appropriately included in data collection related to the treatment environment.

More important than the physical environment is the social environment within which curriculum activity occurs. It has been suggested that for elementary school children the nature of the peer group is more closely related to achievement than is the presence of any sorts of instructional facilities and materials (Coleman et al., 1966). The same results might be expected at the preschool level. Specification of age, ethnic composition, language, size, and stability of the peer group, then, is another way to describe the treatment environment.

The nature of the behavior setting will be established primarily by means of the CIRCUS Educational Environment Questionnaire (CIRCUS test No. 17) and the Planned Activity Check (PLA-Check), with some revisions. Instructional materials and facilities will be assessed independently of any cultural focus they may have, chiefly by means of items 9 through 23 in CIRCUS test No. 17 ("materials, facilities"). These items, to be completed by the teacher, include presence-absence and frequency-of-use information about elements in the physical-social setting potentially relevant to cognitive and social gains. Supplementary information about such environmental elements will be provided in the PLA-Check response form entitled "Teacher's Description of Planned Activities," a data-collection instrument requesting the listing of materials used in connection with each of the day's activities. Neither assessment medium, however, provides adequate information about the use of materials or instructional themes that are specifically relevant to the cultural background of the children attending the class. It is recommended that the PLA-Check teacher's description form be modified so that teachers are asked to mention the use of any culturally focused instructional items or themes. Further, some preliminary work is needed to determine how to classify instructional materials or themes as culturally relevant to given classroom populations so that observers can verify teachers' judgments in this regard. In the discussion of treatment events, it is suggested that such a classification scheme should be made applicable to each activity undertaken during an observation period.

Peer group properties will be ascertained from archival data provided by each Head Start center's records and verified by individual teachers. Information so obtained will include the number of children in the classroom. Additional information to be obtained from archival data will include ethnic composition of the group and group size. The ethnic composition of the group is deemed important in that it will help to describe the peer-bond relationships in the group. This composition will be represented in terms of the proportion of children who are Black, White, Chicano, Puerto Rican, and Native American. The group size will describe the stability of the group in terms of attrition or group turnover.

Treatment Events

The third major set of treatment characteristics are events or activities occurring during the time when children are in the treatment environment. They are the most important indexes of the treatment as process. Specifying the treatment as an ongoing process was regarded by the Rand Classroom Process panel as crucial to the establishment of independent variables capable of differentiating Head Start classes on developmental outcomes. Classroom process here is intended to include an account of regularly scheduled activities as they occur, along with a representation of teacher time invested in planning them and child time spent engaged in them. Of concomitant interest is the degree of control exercised throughout such activities by the teacher. Another aspect of treatment as process is the natural language of the teacher of the classroom. We prefer that this be investigated in a focused study, although it could be investigated in a very small sample of English-speaking classrooms. These aspects of treatment will be considered separately below.

1. *Classroom activity should be measured in all Head Start centers as a representation of the manner in which curriculum design, as translated through teacher goals and attitudes, is eventually incorporated into a set of regular classroom events.* Determining the kinds of activities children undertake in the Head Start class, as well as the proportion of time spent in each, is an objective way of providing information about the educational inputs they are receiving. It is assumed that outcomes are a function of such inputs. For example, we would expect quantitative achievement gains for children in centers that emphasize numerical concepts, but not for children unexposed to these concepts. In the literature reviewed, emphasis on language-learning activities was often significantly related to child outcome measures (although not consistently). Such information needs to be supplemented by data concerning child participation in the scheduled activities. Just as there is a gap between curriculum design and classroom planning, there is often a gap between activity scheduling and active participation. Risley and Cataldo (1974) point out that "the direction and extent of engagement with the physical and social environment appears to be an almost universal indication of the quality of a setting for a people." Their comment calls attention to the need for determining how many of the participants in a group are looking at or physically interacting with materials or people at any moment during a given activity period. Besides providing a qualitative measure of the treatment process, such an assessment would be expected to covary with developmental outcomes. Basic to the operation of most educational endeavors, according to Risley and Cataldo, is the assumption that when a child concentrates

upon a particular activity for a period of time, he gains skill and understanding. Thus, inconsistencies in results related to classroom program emphases might be accounted for by different degrees of participation of children in similar activities in different classrooms.

Degree of child engagement is to be supplemented by the amount of teacher involvement in the activities, assuming that the two are related. In other words, some teachers may select, display, and regulate the use of a wide variety of materials and activities so as to maximize the probability of the children's prolonged engagement with them. As Risley and Cataldo note: "Thus we find that although unformalized in the literature, some teachers do know how to maintain a living environment which will engage children in constructive activities." To assess the teacher's side of the activity schedule, then, it is important to ascertain the amount of time he or she spends planning for each activity, the amount of time allotted to that activity in daily scheduling, the amount of time it actually occupies during the day, and the nature of the teacher's involvement in the activity during the time it occurs. Such information is expected to be related to children's engagement in the activities in question, and it is also useful for determining what distributions of teacher planning and participation time are most productive in this regard.

For the purposes of specifying treatment processes in the manner described, it was decided on the basis of the literature review and panel consultation that only an observation instrument would provide satisfactory measures. Several reviews of observational systems in both classroom and nonclassroom settings are now available (Medley and Mitzell, 1963; Simon and Boyer, 1967 and 1970; and Rosenshine and Furst, 1973). These reviews and the more specialized literature on those observational systems that have been used in the study of young children in nursery and preschool programs provide some 25 documented observational procedures. A discussion of current methods may be found in Wright (1960); Dopyera and Lay (1969); Brandt (1973). There are also newer methods that have not been formally published.

The instruments that were eligible for consideration in the evaluation were those that:

- Included the selected independent variables.
- Were suitable for use in large-scale evaluation.
- Were oriented to classroom activities and processes.
- Focused on the classroom as the primary unit of observation but had sufficient flexibility to permit use of the child as the unit of observation.

At least seven observational systems met the above criteria:

1. Stanford Research Institute (SRI), Classroom Observation Instrument used with HSPV programs, 1969-1972, and Follow-Through (1971-1973).
2. Risley and Cataldo (1973), Planned Activity Check (PLA-Check).
3. Soar (1971), Classroom Process Measures.
4. Medley, Schluck, and Ames (1968), PROSE.
5. Barker and Wright (1949; 1950), Behavior Stream Observation.
6. Brandt (1972), Class Activities Log Sheet.
7. Dopyera (1972), Program Structure Index Procedure.

Additional consideration of the economic feasibility of the observation system, its suitability as an evaluation (rather than research) tool, ease of training, and robustness of measures under diverse circumstances led to a final choice of the Planned Activity Check as the most desirable observation instrument for present purposes.

PLA-Check procedures involve obtaining a schedule of the day's activity periods and a measure of the amount of time spent by children in each activity. The instrument provides a count of the number of children actually engaged with the materials provided by the teacher or teacher aide at a given time in each activity area. It also measures the amount of time the teacher spends planning and participating in the class activities. Since this procedure yields a graphic profile of the group's engagement in each activity and the transition between activities, one can actually tabulate time per activity. The profile illustrates the proportion of time devoted to traditional preschool activities (language, arts, numbers, etc.) and the amount of time engaged in hygiene-related or transitional activities (eating, sleeping, toileting) (see

Risley and Cataldo, 1974). The proportion of time the child and the teacher are each engaged in "on-task" activities can be used as an indicator of the general level of class participation or of the amount of participation in each activity at different times during the day.

PLA-Check does not yield an index of the extent to which children are engaged in culturally focused activities. Although music and dance or art and craft periods are most susceptible to such a focus, some learning materials (e.g., alphabet picture-cards) are often designed to enhance cultural identification. It is suggested that PLA-Check be modified so that each activity period is coded with respect to use of culturally relevant materials or themes. With this exception, the PLA-Check procedures in the standard instrument manual should be followed as described here.

The length of time for observation of each activity area is set according to the amount of detail required in the portrayal of classroom activities, the number of activity areas that must be observed, and the size of the class. For a general picture of the classroom, the developers of the PLA-Check suggest a five-minute interval. Total observation time needed to provide reliable estimates of the teacher and child's degree of engagement in planned activities in the classroom is four observation days in a two-week span. It is advised that different days be sampled across the two weeks--e.g., Monday and Thursday in week one; Tuesday and Friday in week two. The two-week period can be scheduled for November or February, or both. Activity patterns will have stabilized by November and are not yet expected to be disrupted by Christmas preparations. HSPV data indicate that classroom social patterns will have stabilized by February or March. Data from the two time periods (November and February) will probably differ. Although both time samplings would be interesting, the two samples add cost and quantities of data. We recommend one sampling, conducted in February, to keep costs down and to ensure a manageable data base.

One or two observers can be used for each classroom. If two observers are used, inter-rater reliability can be checked throughout the observational period. If only one observer is used, periodic parallel observations must be made by an outside observer, preferably a site supervisor or trainer. In either case, it is necessary that inter-observer agreement be sufficiently high to yield adequate reliability. Risley and Cataldo (1973) state, "two forms, the *Participation Reliability* sheet and the *Attendance Reliability* sheet, provide a means of calculating this percent of agreement between two PLA-Check observers." They go on to say, "Observers who have had some experience with the PLA-Check usually obtain from 80 percent to 98 percent agreement. However, observers who have never used the PLA-Check before often show low agreement the first few times they observe, mainly because of difficulties with timing and synchronization." It is suggested, therefore, that new observers be given several practice sessions in the use of observation techniques before data are collected for use in the evaluation.

2. *Level of control or structure imposed on classroom activities is recommended for assessment in all Head Start centers as an important classroom process variable.* Level of control or structure, the focus of numerous recent studies of educational environments, may be described as the degree to which the teacher or the child has autonomy over the learning activities going on at any given time.

Several researchers have documented the importance of this variable with respect to certain child outcomes. Soar (1971) found the level of "pupil freedom" to be positively related to complex-abstract learning. Stallings et al. (1973) found that in classrooms where teachers allow children to select their own seating and groups part of the time, where many different activities are available, and where there is an assortment of audiovisual and exploratory materials, the children seem to be more independent. Lamb (1965) states that "abstract, complex" teachers (who are less structured and punitive and more flexible and resourceful) increase the child's self-esteem, cooperation, involvement in activities, and achievement. In Rosenshine's (1971) review of several studies of "teacher flexibility," significant relationships were found between cognitive variation and achievement and between environmental enrichment and achievement.

Grannis (1973) has devised a theoretical model that systematically addresses different distributions of control and the consequences flowing from these distributions. He argues that any variable element of the educational environment (e.g., initiation of interactions, physical movement, selection of topics) can be construed as *controlled by the teacher, by the teacher and learner jointly, or by the learner* at any given time or over intervals of time. He also maintains that *congruence* (defined below) in the control of the various elements of the educational environments at a given time contributes to the realization of objectives, and that each of the three levels of control is most appropriate for a different category of educational objectives. Grannis uses the term "congruence" to refer to the condition whereby each level of control has the same hierarchical position as his six stated dimensions (see below).

Grannis' model is regarded as especially useful for specifying the structure of controls in the treatment environment and relating them to educational objectives; the work of other researchers tends to support that framework. For example, the data obtained by Soar and Soar (1973) from the Florida Affective Categories System yielded a factor of "free choice versus structured learning in groups." Data obtained with the Teacher Practices Observation Record yielded a somewhat analogous factor, "teacher-directed activity versus pupil-selected activity." Similar findings are reported with respect to the research carried on by SRI on the examination of relationships between entering ability, instructional processes, and child outcomes. In addition, most of the control variables used by Grannis discriminated quite well between programs and program rationales in the Follow-Through projects. Much of the documentation of these variables as being important determinants of child outcomes is derived from studies of other Follow-Through data or HSPV observational data. These same variables may not discriminate as well between Head Start programs in general, since centers generally do not deliberately adopt specific models. However, our survey of the literature indicates that the available evidence supports the selection of these variables when describing programs even in our broader context.

While level of control exercised in the course of any class of activity is expected to be associated with related cognitive and socioemotional outcomes, correlations are also expected between control variables and general learning goals. It is hypothesized that an educational setting will be effective with respect to child learning to the extent that the control in the setting are congruent. Further, it is hypothesized that each of the three levels of control variables will correlate with a specific type of learning goal, higher correlations obtaining to the extent that controls are congruent. For example, low levels of teacher-pacing of activities should--especially if other sorts of controls are congruent--correlate with Grannis' Type 3 learning described above (major individuation; exploration and application). Thus, level of control is an important process variable completing the specification of the treatment environment in a way coherent with activity and environment descriptions.

There is no existing standardized instrument for assessing level of control as required for the present evaluative purposes. We have considered two methods in considerable detail, both options requiring pilot study during the preparatory year before final decisions are made about measuring level of control. The first and preferred method involves adapting Grannis' operational definitions of control dimensions for use as an observation checklist; the second involves selecting some aspects of the SRI Classroom Observation Instrument and adapting them for measuring control variables with the classroom as the unit of analysis. The Grannis scheme is presented below.

The Grannis theoretical framework isolates and defines six control dimensions,[1] distinguishing three levels of control within each dimension as follows:

- *Task options*. The choice the learner has about the particular task in which he is engaged: (1) prescribed--(no choice) the learner must do this activity now; (2) conditionally prescribed--the child may choose his task from a set of prescribed alternatives; and (3) open choice--an activity that is not prescribed for the learner.

---

[1] Dr. Grannis is in the process of modifying these dimensions (personal communication, August 1974). He is continuing to work with the same basic ideas but is refining the concepts and their definitions. We suggest that he be consulted when these dimensions are operationalized.

- *Prescription of operations*. The degree to which the operations of a task may be determined by the learner: (1) step-by-step sequencing of operations of the learner's activity are completely prescribed; (2) operations are partly prescribed; and (3) operations are not prescribed.
- *Pacing*. The degree to which the learner is free to regulate his work rate or energy output: (1) high teacher pacing--the teacher is continuously present in the learner's setting and continuously makes demands of some sort; (2) medium teacher pacing--the teacher regularly enters and leaves the learner's immediate vicinity; and (3) low teacher pacing--the teacher is absent from the learner's immediate setting, though he may come to the learner occasionally on his own or the learner's initiative.
- *Teacher adaptiveness*. The orientation of the teacher's interactions with the learner to the learner's point of view: (1) the teacher does not alter his behavior with the learner to reflect the point of view of the learner; (2) the teacher alters his behavior, but with the intent of enabling the learner to meet the teacher's criteria for a task or action; and (3) the teacher alters his behavior with the intent of enabling the learner to meet his own individual criteria for a task or action.
- *Materials feedback*. The degree to which the material the learner has in hand confirms the learner's answers to questions the material poses: (1) no feedback, (2) single-answer feedback, and (3) multiple-answer feedback. Emphasis is given primarily to the learner's performance while working with the materials.
- *Interaction among learners*. The degree to which interaction occurs among learners is optional: (1) learner-learner interaction is completely prescribed by the teacher--i.e., prohibited or mandated; (2) interaction among learners is allowed under certain restrictions pertaining to the task; and (3) learner-learner interaction is optional for the child.

According to Grannis, a setting is congruent when one of the following sets of conditions is present:

Level 1: Teacher Control	Level 2: Teacher-Learner Control	Level 3: Learner Control
Prescribed options	Conditionally prescribed options	Open choice on options
Step-by-step sequencing of operations	Operations partly prescribed	Operations not prescribed
High teacher pacing	Medium teacher pacing	Low teacher pacing
Teacher does not alter his behavior (adaptiveness)	Teacher alters his behavior conditionally	Teacher alters his behavior for the learner's advantage
No feedback	Single-answer feedback	Multiple-answer feedback
Interaction among learners is completely prescribed	Interaction among learners is allowed under certain restrictions	Interaction among learners is optional

A setting within a classroom environment is incongruent when the level of control of one or more variables is not consistent with the same level of control of all the variables present. Although several settings can exist in any classroom, a particular class may be categorized as mainly consistent with Level 1, 2, or 3 controls, depending upon the majority of congruent settings. Further, such a description of settings can be correlated to the learning goals (Type 1, 2, 3), suggesting that it is possible to tie educational objectives to the control dimensions of the classroom activities. Grannis' analyses strongly support the hypothesis that congruent distributions of control in learning settings result in more goal-directed behavior by the learners than incongruent distributions of control.

It is suggested that Grannis' six control variables be treated as three-point scale dimensions, any activity being susceptible to rating on all six dimensions. After pilot testing of the reliability of the operational definitions provided above, final versions would be used by observers for making judgments about the extent of control of each activity period observed. The ratings would be combined with the PLA-Check observations in the following way. Although five-minute sampling allows observers sufficient time for making other judgments between activity checks, Risley and Cataldo's data indicate that a great deal of time is occupied in transition from one activity to the next. During each transition, the observer would be requested to rate the immediately preceding activity on each of the six control dimensions.[1] From such data, level of control for each activity can be ascertained as well as congruence of control within and between activities.

3. *Observation of the natural language of the teacher in the classroom is recommended for a focused study, or possibly for a small sample of English-speaking classrooms.* We recommend observing the language environment that the teachers and teacher aides provide for the child.

Tizard et al. (1972) found significant correlations between the language comprehension scores of the children and the amount and quality of adult talk directed at them. Tizard et al. have developed an observational instrument that allows recording of staff verbalizations, staff activity during talk, and whether or not the child responds during staff talk. It does not record the child's speech. We recommend that the measure be administered in a focused study of classrooms. It will be costly to administer and will yield large amounts of data from each classroom. Thus, it is preferable to implement it in a small number of classrooms, either as a focused study or as a restricted subsample of English-speaking classrooms.

CONTROL VARIABLES

The second major set of independent variables includes those that specify the nature of the control condition. The usual interpretation of "treatment group" and "control group" is that members of the treatment group receive the intervention and members of the control group do not. The Head Start treatment has several components. Members of the *control group* receive at least one component of the treatment (care), perhaps more (e.g., education). Thus, an evaluation of Head Start is based on a comparison of the Head Start treatment with treatments that approximate Head Start to a greater or lesser degree.

---

[1] Collection of data on all six control dimensions requires an additional observer. It may be possible to collect data, for example, on three dimensions without adding an additional observer. Cost considerations may necessitate following this latter course of action for the total sample of classrooms. For a smaller sample of classrooms, it may be feasible to collect data on all six control dimensions.

As Chapter 8 indicates, a finding of no difference between the treatment and control groups for a particular site does not imply that Head Start does not have an effect on the children. It may indicate only that most of the control children are enrolled in an effective day care program and that *both* the Head Start and day care programs are benefiting the children. In order to interpret results of these comparisons it is important that we have some idea of how each control child is affected at a site. The outline of control variables below specifies the control condition--i.e., locates the control child within a set of alternatives that are more or less like Head Start. Data on these variables should be collected from the control child's parent at the time of posttest.

I. Type of care during the current Head Start year
    A. Center (institutional) day care
    Period the child was in this type of care
    B. Group (informal) day care
    Period the child was in this type of care
    C. Care in the home
    Period the child was in this type of care

II. Type of care if the child was in the center for ≥ two months
    A. Number of days per week care is offered
    B. Number of hours per day care is offered
    C. Services provided to the child
    A list of the important alternatives should be compiled. Examples are food (breakfast, snacks, lunch); nap; field trips; playing with letters, numbers, words, stories
    D. Number of children in one classroom
    E. Approximate ages of the children in the classroom

F. Ethnic composition of children: Black; White; Chicano; Puerto Rican; Native American

G. Number of adults with each group of children

H. Ethnicity of adults: Black; White; Puerto Rican; Chicano; Native American

I. Educational level of primary caretaker in the group

III. Type of care if child was in group or informal day care for ≥ two months

A. Number of days per week care is offered

B. Number of hours per day care is offered

C. Services provided to the child
A list of important alternatives should be compiled. Since activities in informal day care usually differ from those in center day care, the list compiled for center day care will not be entirely appropriate. For example, we would expect children in informal day care to spend more time in unstructured play

D. Number of children usually in the group

E. Turnover in group membership for the period in which control child was a member: same children; small turnover (≤ 25 percent new membership); moderate turnover (26-74 percent new membership); large turnover (≥ 75 percent new membership)

F. Ages of children in the group

G. Ethnic composition of children: Black; White; Chicano; Puerto Rican; Native American

H. Turnover in primary caretaker: 0; 1; 2; ...; n

I. Relationship of primary caretaker to child: mother; relative (specify relationship); other (specify)

J. Ethnicity of primary caretaker: Black; White; Native American; Chicano; Puerto Rican

IV. Type of care if child was in home care for ≥ two months

A. Relationship of primary caretaker to child: mother; grandmother; sister; brother; father; other relative (specify); family friend; other (specify)

B. Age of primary caretaker

C. Home at which the child stays: own; other (specify)

D. Number of week days the child stays at this home

E. Number of week nights the child stays at this home

F. Number of weekend days the child stays at this home

G. Number of weekend nights the child stays at this home

H. Activities of the child during a usual day and approximate proportion of day devoted to each.
A list of important alternatives should be compiled. Again, it will differ from the major alternatives for center and for informal day care. Examples are watching television, playing with friends, sleeping

I. Number of other children at home with the child

J. Ages of other children at home with the child.

BACKGROUND VARIABLES[1]

We recommend collecting data on the background characteristics of each child; each child's family; teacher (Head Start teacher during the Head Start year; kindergarten or first-grade teacher during the post-Head Start year); institution (Head Start center during the Head Start year; kindergarten or first grade classroom and school during the post-Head Start year); and community, specifically the Head Start catchment area. Table 6-1 lists the proposed battery of background variables; the sample for that battery (i.e., for whom the data are collected); the data source (i.e., from whom the data are collected); and calendar (i.e., when the data are collected for the full-scale evaluation).

The remainder of this chapter lists and briefly discusses the variables for each battery included in Table 6-1.

Child Background Variables

We recommend the following child background variables. These variables are associated with variations in outcomes for the child in the Head Start Planned Variation study or the ETS-Head Start Longitudinal study.

I. Sex of child

II. Ethnicity of child: Black; White; Chicano; Puerto Rican; Native American

III. Age of child in months

IV. Prior preschool experience: Head Start; other preschool; none

---

[1] We would like to thank Virginia Shipman at Educational Testing Service for sharing with us her experience with some of these variables in the *ETS-Head Start Longitudinal Study*.

Table 6-1

BACKGROUND CHARACTERISTICS

Battery	Sample	Data Source	Calendar
1. Child characteristics	Entire sample	Treatment children: Head Start application and attendance records	All variables except attendance: prior to language pretests, Head Start year
		Control children: parents	Attendance records, television questions: at posttest, Head Start year
2. Family characteristics	Entire sample	Treatment children: Head Start records, parents	Language variables: treatment and control children prior to language pretests, Head Start year
		Control children: parents	Other variables: treatment children: at about the time of posttest, Head Start year Control children: at posttest, Head Start year
3. Teacher characteristics	*Head Start year*: All teachers and teacher aides for treatment children	All teachers, teacher aides of Head Start classrooms in sample	Head Start year: optional
	Post Head Start year: All teachers for treatment and control children	All teachers and aides of post-Head Start classrooms in which treatment and control children are enrolled	Post-Head Start year: immediately prior to or during collection of outcome data for children, fall, school year
4. Center characteristics	All centers in the sample	Center personnel	Optional; late winter/early spring, Head Start year suggested
5. Catchment area characteristics	All catchment areas in the sample	Stratified sampling list of centers	Any time during Head Start year
6. Kindergarten or first-grade characteristics	All schools in which treatment or control children are enrolled All classrooms in which treatment or control children are enrolled	School administrators; classrooms	At time of first testing in kindergarten or first grade (fall, post-Head Start year)

V. Proportion of treatment child receives: number of days child attended Head Start/number of days Head Start offered

VI. Child's first language: English; Spanish; bilingual (English and Spanish); other

VII. Television viewing
  A. Hours of television watched per week day
  B. Hours of television watched per weekend day.

Variables VI and VII (child's first language and hours of television viewing) both relate to the child's linguistic background. As Shipman notes, the child of a Spanish-speaking family who watches a great deal of television lives in a more bilingual environment than a child of a Spanish-speaking family who watches little or no television. Knowledge of a child's linguistic status, indicated by child and family background variables, allows us to interpret his or her performance on the Spanish versions of the CIRCUS language measures and to decide whether to administer the English or Spanish versions of those measures to the child. Table 6-1 recommends collecting at least child and family language background data prior to the language pretests so that the tester has a better basis for deciding which version to administer to the Chicano or Puerto Rican child.

Family Background Variables

We recommend the following family background variables:

I. Family structure
  A. Number of siblings
  B. Age of each sibling
  C. Number of adults who are not siblings
  D. Child's primary caretaker: mother; grandmother; other (specify)

II. Language spoken in the home
  A. English only
  B. Spanish only
  C. Spanish and English: 3/4 Spanish, 1/4 English; 1/2 Spanish, 1/2 English; 1/4 Spanish, 3/4 English

D. Other (specify)

III. Language spoken by each member of family

A. English only

B. Spanish only

C. Spanish and English: 3/4 Spanish, 1/4 English; 1/2 Spanish, 1/2 English; 1/4 Spanish, 3/4 English

D. Other (specify)

IV. Mother's education in years

V. Mother's occupational history for the previous four years, compiled by using the coding scheme from the ETS-Head Start Longitudinal study

VI. Father's occupational history for last four years, compiled by using the coding scheme from the ETS-Head Start Longitudinal study

VII. Amount of parents' participation in the community[1]

A. Participation indicated by

1. Number of organizations to which parents belong (e.g., church; Head Start parents' group)

2. Knowledge of local resources available to the family (e.g., library; legal aid; etc.)

B. Lists of relevant alternatives should be developed for A1 and A2.

Teacher and Teacher Aide Background Variables

We recommend the following background variables for Head Start and kindergarten/first-grade teachers and teacher aides:

---

[1] We expect Variable VII to help us explain differences between children within a site and differences between them across sites The variable has a different theoretical status depending on its use. Since participation and knowledge opportunities are fairly constant for all families in a site, variation among families indicates differences in family involvement in the community. We can expect a family that is more involved in the community to be more involved in and reinforcing of the child's experiences, including Head Start. We might also be able to aggregate results by site in order to get an indicator of involvement by site. Site differences can be interpreted as differences in participation and knowledge opportunities, communal *esprit*, or both. These differences become differences in social context for Head Start eligible families, their children, and Head Start centers, and might help to explain some of the inter-site differences in child outcomes that are not attributable to center differences and other differences between families.

- I. Educational level in years
- II. Highest degree held: none; high school certificate; B.A.; M.A.; Ph.D; other
- III. Number of years of paid teaching experience with disadvantaged children: found to be relevant in the System Development Corporation study (1972)
- IV. Sex of teacher
- V. Age of teacher
- VI. Teacher ethnicity: White; Black; Puerto Rican; Native American; Chicano.

Center Background Characteristics

We recommend the following center background characteristics:

- I. Sponsorship of center

  Relevant categories should be developed. Examples are settlement house, Community Action Program, church, Board of Education
- II. Linkages between the center and the community
  - A. We expect differences in treatment effects for children as a function of center involvement in the community. Two alternative causal models for the expected center involvement-child outcome link are:
    - 1. Involved communities cause involved centers and involved parents. Involved parents reinforce the child's Head Start experience and cause children who themselves are more involved in the experience
    - 2. Involved centers involve the parents more extensively in the child's Head Start experience, causing the parents to reinforce that experience more consistently
  - B. Indicators of linkages between the center and community should be developed. The number of different activities for which the Head Start center is used and frequency of occurrence of each activity type might indicate community connections. In this case decisions have to be made about

1. Definition of activity--for example, providing information on community resources (e.g., medical help) should be considered an activity
2. Procurement of data on types and frequency of activities--one source of data might be mimeographed event calendars

III. Proportion of center staff who reside in the catchment area[1]

IV. Physical environment (included in treatment variables above)

V. Money spent per child.[2]

Site (Catchment Area) Variables

We recommend the following site background variables:

I. Metropolitan/nonmetropolitan: urbanized; less urbanized; sparsely populated. These terms are defined in Chapter 7 or p. 303.

II. Region--We recommend the census divisions of the U.S. Bureau of the Census: New England; Middle Atlantic; East North Central; West North Central; East South Central; West South Central; South Atlantic; Mountain; and Pacific. Operational definitions of these divisions are provided in U.S. Bureau of the Census (1970). We especially recommend examining outcome data in terms of the *intersection* of variables I and II.

There are several variables, initially promising, that were discarded for theoretical or measurement reasons. For example, the ethnic and SES composition of a community represents a structure of social comparisons for families within that community. We know that the social comparisons available to people make a difference. However, the effect of those comparisons seems to vary. We also do not know if the psychological community for Head Start parents is the same as the catchment area. In sum, ethnic and SES composition of site seems less helpful than ethnic composition of the Head Start classroom.

---

[1] Variation in proportion of staff who derive from the catchment area may affect parental acceptance of the center and parental reinforcement of the child's Head Start experience. When most staff reside outside of the catchment area, this variable may prove to indicate a center unconnected with the community.

[2] We have two major reservations about this variable. First, Coleman et al. (1966) has shown that at least for schools, financial inputs do not affect outcomes. Second, the economy for many centers is partly a barter economy. It thus becomes very difficult to measure resource input reliably.

A second variable of this sort is political activism of the community. We sense that psychological and social life is very different for individuals resident in politically active and participatory, as opposed to apathetic, communities. We might be able to use aggregation of parental participation by site as an indicator of this variable. (See family background variables, above.) Again, however, it is not clear that the activist unit for Head Start families is coincident with catchment areas. We also would expect this activism to register on the child more directly through the activist quality of the center and the parents. If the children were adolescents, we would expect the dynamism of the community to affect the individual directly. However, Head Start children range from 3 to 5 years in age. They are not really members of the community yet, except as a function of membership in family, neighborhood, or the Head Start center. The activist quality of these units is more relevant than that of the general community. Valid, reliable, and inexpensive measurement is a third problem. Indicators of this dimension in one community are often not valid indicators in another. Good participant observation yields trustworthy data, but the cost of collecting the data exceeds what independent information the effort could probably yield. For these reasons, we do not recommend trying to scale communities on political involvement.

Kindergarten and First-Grade Variables

We are interested in properties of kindergarten and first-grade classrooms and schools in order to place socioemotional outcomes, most of which cannot be measured until the post-Head Start year, in their measurement context. We recommend that the contractor reconsider the following variables in light of the pilot-run experience:

I. SES composition of kindergarten/school

II. SES composition of classroom

III. Ethnic composition of classroom: proportion of Black; White; Puerto Rican; Native American; Chicano

IV. Location of the kindergarten or elementary school in or outside of the Head Start catchment area

V. Proportion of children in classroom who attended Head Start

VI. Curriculum model/plan for classroom

A. Consistency--a Head Start child may do less well than the control child in post-Head Start classrooms, if the Head Start and kindergarten and first-grade classrooms use radically different curricula models.

B. Measurement--we recommend the same modification of the Weikart classificatory scheme used in measurement of the Head Start treatment variables.

Chapter 7

BASIC EVALUATION DESIGN

INTRODUCTION 265

SAMPLING RATIONALE 268

SELECTION OF TREATMENTS 271
  Sampling Procedure for Selecting Treatments 271
  Restrictions on the Sampling List of Centers 274
  Selection of Random Subsample and Special Subsamples 274

SAMPLE OF CHILDREN 274
  Selection from the Eligible Set 275
    Volunteers for Treatment 275
    Handicapped Children 276
    Unequal Treatment Period 277
    Children of Migrant Workers 278
  Selection for Treatment and Control Groups 279
    Alternative Designs 279
    Importance of Random Assignment 285
    Implications of Random Assignment for Types of Centers ... 289
    Recommendations for Conducting Random Assignment 291

STRATIFICATION DIMENSIONS FOR TREATMENTS AND CHILDREN 294
  Rationale and Criteria 294
  Choice of Treatment Dimensions 296
  Choice of Children Dimensions 298
    Cultural Dimension 300
    Metropolitan/Sparseness and Regional Dimensions 302

SAMPLE SIZE 308
  Sample Size Decisions 308
  Criteria for Sample Size Decisions 309
    Relationship between Sample Size and Precision 309
    Optimum Allocation 313
  Recommended Sample Sizes 315

Chapter 7

BASIC EVALUATION DESIGN[1]

INTRODUCTION

As indicated in Chapter 2, the basic evaluation is addressed to four questions:

1. What are the social competence effects of Head Start for members of the eligible population who receive the treatment, relative to members of that population who do not?
2. What are the social competence effects of Head Start for eligible children from different cultural groups who receive the treatment, relative to eligible children from those same groups who do not?
3. What are the social competence effects of Head Start for eligible children within each cultural group who receive the treatment and who differ in entry characteristics, as indicated by pretests and other background characteristics?
4. Are there any indications that variations in treatment produce variations in social competence outcomes for children who receive the treatment?

Table 7-1 presents the structure for the data collection for all four questions. It assumes random sampling of sites and random assignment of eligible volunteers within site between T and -T conditions. In analysis of variance terms the design is known as a randomized block, partial hierarchical design. However, for reasons discussed below and in more detail in Chapter 8, the analysis model associated with that design is not appropriate for any one of the questions.

---

[1] We want to thank a number of persons for their thoughtful contributions to this design: Robert Boruch, Northwestern University; John Butler, Harvard University; Anthony Bryk, Huron Institute; Andrew Porter, National Institute of Education; Lee Sechrest, Florida State University; and Pierce Barker, Stephen Carroll, Carl Morris, Peter Morrison, William Rogers, and John Wirt of The Rand Corporation.

Table 7-1

STRUCTURE OF THE DATA COLLECTION

Stratum	Site	Treatment Levels	
		T	-T
Stratum_1	Site_{11}	$O_{1111}, O_{1112}, \ldots O_{111n}$	$O_{2111}, O_{2112}, \ldots O_{211n}$
	Site_{12}	$O_{1121}, O_{1122}, \ldots O_{112n}$	$O_{2121}, O_{2122}, \ldots O_{212n}$
	$\vdots$	$\vdots$	$\vdots$
	Site_{1k}	$O_{11k1}, O_{11k2}, \ldots O_{11kn}$	$O_{21k1}, O_{21k2}, \ldots O_{21kn}$
Stratum_2	Site_{21}	$O_{1211}, O_{1212}, \ldots O_{121n}$	$O_{2211}, O_{2212}, \ldots O_{221n}$
	Site_{22}	$O_{1221}, O_{1222}, \ldots O_{122n}$	$O_{2221}, O_{2222}, \ldots O_{222n}$
	$\vdots$	$\vdots$	$\vdots$
	Site_{2k}	$O_{12k1}, O_{12k2}, \ldots O_{12kn}$	$O_{22k1}, O_{22k2}, \ldots O_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Stratum_j	Site_{j1}	$O_{1j11}, O_{1j12}, \ldots O_{1j1n}$	$O_{2j11}, O_{2j12}, \ldots O_{2j1n}$
	Site_{j2}	$O_{1j21}, O_{1j22}, \ldots O_{1j2n}$	$O_{2j21}, O_{2j22}, \ldots O_{2j2n}$
	$\vdots$	$\vdots$	$\vdots$
	Site_{jk}	$O_{1jk1}, O_{1jk2}, \ldots O_{1jkn}$	$O_{2jk1}, O_{2jk2}, \ldots O_{2jkn}$

A *site* is defined as the catchment area for a Head Start center. The *strata* represent *cultural* distinctions among the total set of eligible children and are discussed in detail in the section on stratifications of children. The treatment factor has two levels, Head Start (T) and not-Head Start (-T). *Head Start* is defined as whatever goes on in a Head Start classroom. *Not-Head Start* is defined as whatever happens to eligible children not enrolled in Head Start during the hours in which Head Start classes meet. This can vary from staying at home to attending a formal day care center.

It is important to understand the constraints the Head Start program imposes on a design. First, we expect considerable variation *within* each treatment level--i.e., within Head Start (T) and not-Head Start (-T). "Local options" are *mandated* as part of the Head Start program; thus, T should vary from site to site. -T inevitably varies between sites, depending on the availability of preschool or child care alternatives to Head Start. In this situation treatment levels are almost certainly confounded with sites. We then have no way of obtaining an unbiased estimate of treatment effects across sites.

A standard solution to this problem is to (1) type T and -T experiences into variations that are expected to predict to differences in outcomes for children, and (2) randomly assign variations to sites. Chapter 6 proposes variables from which such a typology might be developed. However, we do not know if these variables will successfully differentiate within T and -T experiences. Even if such a typology can be developed, we cannot randomly assign variations to sites. Such a move violates the tenet of community flexibility.

We recommend a solution, discussed in more detail in Chapter 8. If eligible volunteers are randomly assigned to T and -T conditions, we can obtain an unbiased estimate of within-site effects. The estimate of overall Head Start effect is then based on the aggregation of site-specific results.

Another constraint is that children cannot be randomly assigned to sites. We can then expect confounding of site effects and child background characteristics. This is not a problem if we want to think of site and child background variables as blocking variables--i.e., as ways of disaggregating the total sample of children into maximally homogeneous groups. In this case confounding is efficient. It is a problem if we want to obtain unbiased estimates of the contributions of site (community) and child background characteristics to children's outcomes.

Further, sites are nested in and cannot be randomly assigned to levels of the stratum, or cultural, factor. Since rates are assigned to a cultural category as instances of that category, sites and cultural categories are completely confounded. Thus, we cannot estimate the independent contributions of site and cultural category to outcomes for children. We do not see this as a problem. The cultural factor is a "sampling factor": It is the basis for constructing the stratified sample required for question 2.

The design and the special analysis model (see Chapter 8) yield estimates for questions 1 and 2. The design cannot disentangle variations within T and -T from site effects and child background characteristics from site effects. Thus, it is not structured to yield *a priori* estimates for questions 3 and 4. These questions can still be evaluated, but in exploratory and *ex post facto* ways.

The remainder of this chapter presents the design for the basic evaluation. The chapter is organized into five major parts: sampling rationale, selection of treatments, selection of children for treatment and control conditions, stratification dimensions for treatments and children, and sample size.

SAMPLING RATIONALE

We want to estimate the effect of a particular treatment (Head Start) on children eligible for the treatment. As indicated in the Introduction, the treatment can be defined as everything that occurs in Head Start classrooms. Since there are approximately 26,000 classrooms, there are approximately 26,000 elements in the treatment set. Head Start serves about 400,000 children. This number is estimated to be 10 to 30 percent of the total number of eligible children. In other words, there are somewhere between 1.33 and 4 million Head Start-eligible children.

Sampling is relevant under two conditions: when there are too many items in the population to evaluate each individual, and when one desires to make statements about the total population. Both conditions apply to the evaluation of Head Start. For example, consider the first condition: There are approximately 15 children per Head Start classroom. If we assume no geographical constraints, there are theoretically an infinite number of ways in which to draw 15 items from a set of 1.33 to 4 million. In actuality, however, geographical constraints obtain. We are not concerned with the effects of Head Start on 15 children drawn from 15 different points in the United States and exposed to the treatment at a 16th point. The combinations of practical interest are all the possible combinations between each of the 26,000 classrooms and the eligible children in the catchment area for that classroom. In sparsely populated areas there may be only one possible combination: The Head Start center has only one classroom and the total number of eligible children in the catchment area is 15. Thus, there is only one way to choose a set of 15 children and one way to assign the set to a treatment instance. In this case, the number of possible combinations of treatment and children is 1 × 1, or 1. In a more densely populated area the catchment area may contain 300 eligible children and six Head Start classrooms. The number of possible different combinations is

$$\binom{300}{15} \times 6, \quad \text{or} \quad \left(7.7101 \times 10^{24}\right)\left(6\right).$$

The probability that we will observe any particular sample of 15 in combination with any particular one of the six treatment instances is

$$\frac{1}{\binom{300}{15} \times 6} \quad \text{or} \quad \frac{1}{\left(7.7101 \times 10^{24}\right)\left(6\right)}.$$

If we draw samples of size 15 from the set, the number of ways we can fill the six classrooms is

$$\binom{300}{15,\ 15,\ 15,\ 15,\ 15,\ 15,\ 210} \times 6!$$

The implications of these numbers are (1) for any one site there can be a large number of combinations of treatment and subjects, even if the relevant set of children is restricted to those who volunteer for the treatment; and (2) for the total set of treatment instances (approximately 26,000) there is an infinite number of combinations of treatment and subjects. To evaluate the effect of the treatment on eligible children, it becomes necessary to select from these combinations.

The second condition under which sampling is relevant--desire to make statements about the total eligible population--also pertains to the Head Start case. Question 1 about effects of Head Start on children is a question about the total population of combinations of treatment and catchment area. Thus, not only is it necessary to sample from these combinations, it is also necessary that the selection be made so as to allow us to treat sample data as estimates of population parameters. We use statistical sampling techniques because they provide rules for selecting items from a population in a way that allows results for the items to be generalized to the population.

Campbell and Stanley (1963) make a distinction between the *internal validity* and *external validity* of experimental designs. An internally valid design excludes explanations of the results other than the explanation of treatment effect. For an externally valid design, results can be generalized legitimately beyond the experimental situation.

The procedure for sampling sites and classrooms within centers affects the generalizability of site-specific conclusions to the total population of sites. The procedure for sampling children from within a site affects the generalizability of conclusions to all children within all sites. The procedure for assigning selected children within a site to treatment and control conditions affects the internal validity of the conclusions. *The internal validity of the conclusions is logically prior to their generalization.* A set of sites may represent a valid frame for generalizing conclusions from specific sites; however, if the conclusions for individual sites are questionable, a valid generalization framework does not help them. It only extends the problems to the population of sites.

Obviously, an optimal design is internally and externally valid. However, under field conditions tradeoffs can develop between the two. In that case we strongly recommend that internal validity take precedence over external validity. The rationale for this recommendation is developed below in the section entitled "Sample of Children."

SELECTION OF TREATMENTS

This section is organized into three parts: sampling procedures for selecting treatments; restrictions on the sampling list; and selections of random subsamples and special subsamples.

Sampling Procedure for Selecting Treatments

To obtain a sample of treatments, we have two basic choices: simple random sampling or cluster (multistage) sampling. Either alternative can be stratified or unstratified. In the simple random sampling case, classrooms are randomly sampled from a complete list of Head Start classrooms. There are two stages in the cluster sampling case: (1) Head Start centers are randomly sampled from a complete list of Head Start centers, and (2) for each center in the first sample with two or more classrooms, Head Start classrooms are randomly sampled from a complete list of classrooms for that center.

We recommend two-stage cluster sampling for the national evaluation. The argument for this choice is as follows. As indicated in the Introduction, we recommend that each site be able to stand as a separate experiment. This implies that the sample size at a site be sufficient to protect the power of statistical tests applied to data collected at that site. We can assume some level of attrition (replacements with less than a complete course of treatment). Certainly the attrition probability varies by site--e.g., urban versus rural probabilities are different. However, assuming 15 children per classroom, even under conditions of low attrition the number of usable data points from the average classroom will be below 15. The relationship between sample size and statistical power varies as a function of other variables. However, as shown in detail below, statistical power is small for small sample sizes (e.g., 10 or 15), unless the ratio of difference between treatment group means to experimental error is large. Pilot test data can be analyzed to estimate this ratio for different measures. If (1) the difference between treatment means is ≥ 1.5 times experimental error for each measure in the basic battery, and (2) maximum predicted attrition will not pull classroom size below ten, then one classroom per center will yield enough data points for the analysis. In that case, treatment instances can be randomly sampled from a complete list of Head Start classrooms.

The probability is that the ratio of the difference in means to experimental error will not be large for all measures. It then becomes necessary to enlarge the number of data points per site by obtaining data on children from two or more classrooms per site. This strategy requires a cluster sampling procedure.

We can select a sample of centers from a list of centers either by simple random sampling or by systematic sampling. In simple random sampling, each draw is independent and random. In systematic sampling, the first item is selected randomly. That draw determines all subsequent draws. For example, there are approximately 8000 to 9000 centers, assume 8000. For purposes of this example, assume that we do not want to stratify the list of centers. If we want a total sample size of 160 centers, this is a sampling fraction of 1/50. Assume that a number is randomly selected between 1 and 50, say, 37. The membership of the sample is now determined. The second item is 37 + 50 = 87; the third, 87 + 50 = 137; etc. The total sample is defined by:

$$S = r + (s - 1)\delta; \; s = 0, 1, 2, \ldots, \frac{1}{\delta} \cdot N,$$

where S = total sample,
N = population,
r = random start number,
δ = sampling interval.

Systematic sampling distributes the sample more evenly over the listed population. To the extent that the population is variable and the variability tends to cluster, an evenly distributed sample is more accurate. There are several possible disadvantages of systematic sampling. One is that populations can have either periodic or linear systematic variation. In the periodic case, the sample is biased if the interval between successive units in the systematic sample coincides with the wave length or its multiple. In the linear case, the sample is biased if the problem is to estimate the average value of a phenomenon that has linear properties. For example, assume that the problem is to estimate density of automobiles within a six-mile radius of the city center for 50 cities. If the random start measurement point for city 1 is 0.5 miles, the density within the area 0 to 0.5 miles from city center will not enter the estimate of mean automobile density. One way to protect against systematic variation is to combine simple random and systematic sampling techniques. In other words, the population is divided into groups of sampling fraction size. A new random start number is then chosen for every other or every third group.

As discussed below, we recommend stratifying the sample into culturally distinct groups. *If analysis of the pilot test data reveals variability within strata of the sample, we recommend systematic sampling of the center population.*[1] This move should produce a more accurate within-stratum sample. It is difficult to think of any periodic variation that could occur within a stratum of the population of Head Start centers. Linear variations are conceivable. Centers may be on lists in some rough order of size, urbanness, quality, etc. If these circumstances are thought to pertain, it is advisable to use simple random sampling or to choose frequent new random starts.

We recommend simple random sampling for choosing Head Start classrooms from within a center. Within-center variation should be small. If there is systematic linearity in classroom lists, systematic sampling could produce extremely biased samples. We assume that on the average there are three classrooms per center. Each time the sampling fraction for a center does not divide the classroom population into equal intervals, a significant portion of the total population at the end of the list drops off. For example, if there are five classrooms in a center and the sampling fraction is two-fifths, a classroom is selected from the first and second intervals of two. The fifth classroom on the list "drops off." If the classrooms are randomly arranged on the list, there is no problem. However, if they tend to get ordered in some systematic way, systematic sampling would produce a disastrously biased sample.

---

[1] If the center population is stratified, systematic sampling occurs within strata. The random start number is randomly chosen for *each* stratum.

Restrictions on the Sampling List of Centers

Although we have limited knowledge about theoretically relevant variation among Head Start treatments, the recently introduced I and I (Improvement and Innovation) program allows different options, some involving two-day per week and other changes in amount of treatment. There is reason to think that variations in amount of treatment make a difference (e.g., the Westinghouse results (1969) for summer versus full-year programs). In the belief that most centers still offer the standard five-day per week treatment and in the interests of holding at least one source of treatment variation constant, *we strongly recommend that the basic evaluation be restricted to the more standard five-day per week treatments, either half-day or full-day.*

Selection of Random Subsample and Special Subsamples

The preceding chapters have referred to "subsamples" and "special subsamples." A subsample is recommended in cases where a measure is too expensive to administer at all sites in the total sample. A special subsample is recommended when a measure is inappropriate for all strata.

A subsample is defined as a randomly and proportionately selected subset of the total stratified sample. A special subsample is defined as the total, or a proportional subset of the total, sample for *selected strata*. Thus, a subsample is selected by randomly drawing an equal proportion of centers from the total sample of centers for each stratum. A special subsample is drawn either by taking all centers in selected strata, or by randomly selecting an equal proportion of centers from the total sample of centers for those strata.

SAMPLE OF CHILDREN

The previous section states the basis for selecting sites (i.e., Head Start centers or catchment areas). This section addresses the problems of how children in a given catchment area are sampled from the eligible set in that area and are assigned to either the treatment group or the control group.

Selection from the Eligible Set

We suggest excluding four groups from the set of eligibles to whom results of the evaluation can be generalized: nonvolunteers, handicapped children, children who receive unequal periods of treatment, and migrant children.

Volunteers for Treatment. Head Start eligibility is defined in the federal legislation for Head Start. If we are able to list all eligibles in a sampled catchment area, it is possible to select children randomly for treatment and control conditions. In this case, the cross-site results of the evaluation are legitimately generalized to the set of treatment instances *and* to the set of eligible children.

The problem with random selection from the set of eligibles for a voluntary program is acceptance of the treatment. The usual Head Start program is five-days per week for approximately eight months. Randomly selected individuals are generally willing to accept treatment of very limited duration (e.g., responding to a questionnaire). However, a treatment of the length of Head Start requires substantial commitment. A behavioral indication of commitment is voluntary request for the program. If the evaluation is limited to volunteers for the program, the results can be generalized only to the set of children whose families request the program. One solution to the problem is to work with three groups of subjects: those assigned to the control groups, those invited to participate in the program who accept, and those invited to participate in the program who refuse. The difficulty with this solution is that the control group consists of individuals who would have accepted the treatment if invited to participate, and those who would not accept. There is no reliable way to discriminate between these two groups. Thus, the treatment group must be defined as both groups of invitees, whether or not they accept the treatment. The treatment group then has at least some individuals who do not receive the treatment. Unless treatment effects are very strong, the dilution of the treatment group may obscure genuine effects.

Restricting the generalizability of results to volunteers has costs to the extent that Head Start is expected to become a compulsory program or one that serves a much larger proportion of eligibles. Neither e[illegible]lity seems likely at this time. *We recommend that members of the control and treatment groups be self-selected, not randomly selected.*[1]

Handicapped Children. A handicapped child is by definition an outlier on the dependent variables. Head Start is mandated to serve handicapped children. However, for some conditions and on some dependent variables these children come from a different population than do other eligibles in the catchment area. If the T and -T groups have the same number and *type* of handicapped children, the statistical problem is solved. If we want the effect estimate to be the effect for all children, handicapped or not, there is also no substantive problem with aggregating scores of handicapped and nonhandicapped children. However, handicapped children are apt to have special problems; treatment effects are apt to be different for them than for the nonhandicapped child. If we want to understand effects for the handicapped as a distinct group, aggregating the data for handicapped and nonhandicapped children yields a distorted estimate for both. If, for example, a classroom has an average of 10 percent handicapped children, in two classrooms there are approximately three handicapped children for a total of 30 children. This means that even if scores for the handicapped are outliers, their contribution to the effect estimate is small. Thus, the effect estimate does not represent the effect for handicapped children. At the same time their scores disproportionately affect the effect estimate for nonhandicapped children. Thus, the estimate does not accurately reflect effect for nonhandicapped children.

---

[1] Even if random sampling is eliminated from the design, there are ways to increase the representativeness of those who select the treatment. Individuals may fail to select treatment for reasons other than lack of interest. They may not know the program exists; they may not know they are eligible for it; they may not know how to enroll. Head Start programs are supposed to have recruiting procedures designed to eliminate these barriers to participation. Therefore, one way to maximize the representativeness of the subjects who select the treatment is to be certain that sites in the sample use those procedures.

If recruiting procedures of this sort are not generally distributed across Head Start centers, evaluating programs with these procedures means evaluating programs atypical on at least this dimension. *A priori*, one would expect atypicality here to be associated with atypicality on other, theoretically relevant, variables (e.g., community participation).

There are two problems with applying the criterion. One problem is deciding which child is handicapped. Does the evaluation use the Head Start decisions on which children are handicapped? Studies on the handicapped (e.g., Kakalik et al., 1973) have shown that there is considerable error in categorizing children as handicapped or not handicapped, and according to type of handicap. However, medical pretests for the evaluation are scheduled as optional, and those for the complete sample are not structured to detect all types of handicaps. They detect hearing and vision problems, etc. Unless special money is included in the evaluation to screen handicapped children for proper diagnosis, the best, although not ideal, solution seems to be to use the Head Start center categorizations of children.

The other problem is deciding which handicaps are serious enough to produce clearly different problems and response possibilities for the child. Some handicapping conditions are mild. In these cases the child is not substantially different at entry, and his or her response possibilities are not particularly different from those of nonhandicapped children. If pretest scores for handicapped children do not clearly discriminate them from nonhandicapped children, the children might be included in the sample.

Our preferred solution to evaluating the effects of Head Start on handicapped children is by means of a focused study. Such a possibility is discussed in detail in Chapter 10.

Unequal Treatment Period. Attrition from the T group occurs across the Head Start year. Thus, replacements have less than the full treatment. In other cases, children receive a second year of Head Start. It is reasonable to expect Head Start effects to vary with length of treatment. On this basis we recommend that children who have more or less than one round of treatment be excluded from analyses of Head Start effects. We suggest using the date of pretests as the cutoff point for late entrants. Pretest dates are selected to measure the children before the treatment can be expected to have an effect.

It would be useful to know the effects of shorter and longer treatments, and some independent variables are selected to yield this information (see Chapter 6). However, estimating effects of variation in length of treatment from scores of children who drop out, come in late as replacements, or stay in longer have several problems. Children who drop out or stay in longer than one treatment round are plausibly different from other children. Late entrants have a different learning situation than initial entrants. They are usually faced with established cliques and performance situations unfamiliar to them, but familiar to other members of the group. Evaluating the effect of variation in length of treatment is preferably done by a small experiment, where length of treatment is systematically varied and children are randomly assigned to the different conditions.

Children of Migrant Workers. There are five streams of migrant workers in the United States. One in South Texas is composed of Chicanos, and 55 percent of the migrant children served by Head Start belong to this group. This stream spends four to six months in Southern Texas. It moves to the Panhandle of North Texas or to an area bounded at the East by New York and at the West by Washington State. The children enter Head Start centers at their new places of residence.

The second stream is Black and based in Florida. It leaves the state only six weeks to three months out of the year, moving up and down the state for the remainder of the year. A Head Start program that allows children to enter centers along the migrant route is in the planning stages.

The third stream is primarily Chicano and based in Arizona and California. It is fairly stable, moving from north to south in California and back and forth between Arizona and California.

A small Native American stream of Navahos based in Arizona and Kikapoos based in Oklahoma moves into Utah and Idaho. The Utah Migrant Council works with those Indian children who migrate into the area.

The fifth stream is a group of Mexican citizens that moves into New Mexico and does not have legal access to Head Start.

There are approximately 7600 migrant children in Head Start. For the basic evaluation, the population of children could be stratified to include a migrant category. However, we recommend that the evaluation of Head Start effects on migrant children be treated as a focused study. In other words, *we do not recommend including migrant children in the* *basic evaluation*. The reason is as follows. The life situation for migrant children is different from that for children of their same culture group in their base area. We do not know whether that difference is important. If they complete the Head Start program, it is probably because the program is set up so that children can enter centers en route. These centers are probably different in certain ways from centers for nonmigrant children. It is not known whether these differences are important. Thus, there is reason to treat migrant children as a separate stratum and to consider Head Start centers that deal with mobile children as a separate treatment.

Three of the four streams are culturally different from each other. In other words, migrant children have mobility in common. Relative to nonmigrant members of their culture group, they have less access to services. However, in other respects they are more like same-culture, nonmigrant children. The implication for stratification is to set up three categories: migrant Black, migrant Chicano, and migrant Native American. We expect that it would be very difficult to get large enough samples of migrant Head Start treatments to estimate effects of Head Start for each of these three categories. Similarly, given the special nature of children and treatment, we think it would be preferable to design special studies to shed light on the special needs of these children and to estimate the effects of special treatment on them.

Selection for Treatment and Control Groups

Alternative Designs. There are a number of alternative designs by which children can be placed into treatment (T) and control (-T) groups. We consider eight alternatives:

1. Two-group design
   a. Random assignment of volunteers to each group
   b. Pretests recommended, not necessary
2. Two-group design
   a. Tie-breaking random assignment of volunteers to each group
   b. Pretests

3. Two-group design
    a. Regression-discontinuity assignment of volunteers to each group
    b. Pretests

4. Two-group design
    a. Assignment of volunteers to the T group
    b. Assignment of eligible nonvolunteers from the same catchment area to the -T group
    c. Pretests

5. Two-group design
    a. Assignment of volunteers to the T group
    b. Assignment of eligible nonvolunteers from a different or a noncatchment area to the -T group
    c. Pretests

6. Two-group design
    a. Assignment of volunteers to the T group
    b. Assignment of eligible nonvolunteers from the same catchment area to the -T group
    c. No pretests

7. One-group design
    a. Assignment of volunteers to the T group
    b. No -T group
    c. Pretests

8. Value-added design
    a. Assignment of volunteers to the T group
    b. No -T group
    c. Pretests
    d. Growth curves.

These designs vary substantially in their internal validity.[1] Of the eight designs, Design 1 is the only one that protects the interpretability of the results for the *total* set of volunteers. There are infinite pretreatment differences between children that can cause differences on outcome measures. If sufficient numbers of volunteers are randomly assigned to the T and -T groups, in general it can be assumed that these differences are equally distributed between the two groups before treatment. Thus, post treatment differences between the T and -T groups on outcome measures can be attributed to effect of the treatment.

---

[1] No design excludes this explanation of T and -T posttest differences: T sensitizes children to an external opportunity--e.g., "Sesame Street"--which causes the T and -T posttest differences. This explanation is evaluated by observing the effects of treatment variations, not by varying the method of selecting children for the treatment.

Designs 2 and 3 are similar. They are appropriate to ameliorative interventions where the intervention is a scarce resource; it is allocated to the most needy among the eligibles; and "most needy" is defined by a quantifiable eligibility criterion that is correlated with the dependent variables. These designs are appropriate to the Head Start evaluation in catchment areas where demand for Head Start exceeds supply. In both designs, children are ranked according to the need criterion. In Design 2, children who fall into the interval containing the cutoff point are randomly assigned between T and -T groups. Any difference (discontinuity) between T and -T children from that interval can be attributed to the treatment. Design 3 is a quasi-experimental version of Design 2; it involves no random assignment of any group. Data from each dependent variable are plotted against the eligibility criterion. The plot is used to extrapolate to the results of a hypothetical tie-breaking experiment at the cutting point interval, rather than to estimate effects at all eligibility levels.

There are several disadvantages common to Designs 2 and 3. First, both require an eligibility criterion whose cutoff point interval is associated with different values of the dependent variable, depending on presence or absence of treatment. Second, both require a public and reliable basis for ranking children on eligibility. Third, even if the cutoff point interval is broad--and this creates problems for Design 3--only a small number of cases will be relevant to the estimate of treatment effects. Even if there are treatment effects, they may be difficult to detect in a small sample. Fourth, both designs explore treatment effects for only a narrow range of eligibility.

Design 3 has additional problems. For example, there has to be an assumption about the functional relationship between the dependent variable and the eligibility criterion. If the relationship is sigmoid (S-curved), subjects just below the cutoff point may be higher on the dependent variable than those beyond the cutoff point, independent of treatment.

Design 4 is a fairly weak quasi-experimental design. It does not allow random assignment of subjects to T and -T groups or random assignment of treatment to preexisting groups. Thus, results on the dependent variables can be a function of pretreatment differences associated with volunteering for treatment, not a function of the treatment. Pretest scores and scores on the other child characteristics can be used as covariates to equate the T and -T groups. However, the less reliable the covariate, the less well it adjusts differences between the groups (Lord, 1960, 1967, 1969; Porter, 1967). Even if the covariates are highly reliable, there is still the problem of relevant unmeasured properties.

Design 5 is a design that can be used when all eligibles in a catchment area are enrolled in Head Start. This can occur when the eligible population is very small (e.g., in sparsely populated areas) or when the supply of Head Start services is large. Obtaining a -T group of eligibles outside of the catchment area of the T group allows a two-group design. However, it has the problems of Design 4 and the additional problems of greater noncomparability between T and -T children, as a function of different community.

Design 6 is an *ex post facto* design. It has the problems of Design 4 and the additional problem of having no pretest data for the T and -T groups. Thus, it is not possible to know whether results are attributable to the treatment, pretreatment differences on the dependent variables, or unmeasured differences associated with volunteering for the treatment. Essentially, the data are uninterpretable.

Design 7 also yields uninterpretable data, but for different reasons. Pretests locate each child on the dependent variable before treatment. Thus, we can tell if the child changes during the treatment period. However, the absence of a comparison group means that any difference between pretest and posttest can be attributed to such circumstances as maturation, rather than to the treatment.

Although it is superficially similar to Design 7, Design 8 is a promising design when *growth curves* are available for untreated Head Start volunteers from different strata for the specific outcomes and measures selected for this evaluation. A growth curve for untreated subjects is a criterion against which to measure the performance of treated subjects. In other words, a physical control group is not necessary because the growth curve operates as a control group.

As indicated, to be able to use a growth curve design, it is necessary to obtain highly specific curves (i.e., empirical "time series") for: (1) total age range of Head Start eligibles; (2) eligibles from each stratum who are untreated; (3) eligibles who volunteer for the treatment; and (4) each outcome and measure of interest. Obviously, these requirements for growth curves are analogous to the requirements in experimental design for comparability of a control group to a treatment group.

Curves of this sort have been constructed for Head Start data. The HSPV study uses a quasi-experimental design with dubiously comparable controls. In the *analysis* of these data Smith (1973) and Weisberg (1973) tried to estimate treatment effect by using the treatment children as their own controls. In a strategy that came to be known as "value-added analysis" they constructed *ex post facto* growth curves for each measure of interest from the pretest scores of children who volunteered for the treatment. They assumed that *pretest* scores for children aged j months represented what the scores for children aged j months at *posttest* should be *in the absence of treatment*. Any difference between expected and observed posttest scores could then be attributed to treatment effect.

Bryk and Weisberg (1974) elucidate the statistical theory behind value-added analysis. They also compare a value-added analysis of HSPV data with a standard analysis of covariance of the same data. Estimation of treatment effects are consistently *smaller* for the value-added analysis than for analysis of covariance (ANCOVA). Bryk and Weisberg argue convincingly that the value-added analysis yields less biased estimates of treatment effect for the data than does ANCOVA.

Thus far, the value-added concept has been used only to analyze data that, given the study design (Design 5), would ordinarily have been analyzed by ANCOVA techniques. However, there is no reason why the value-added concept cannot be used as the basis for a design. In the Head Start evaluation case, growth curves would be established by sampling centers so there would be an adequate number of sampling points for all ages represented in the Head Start program for each stratum. This would require over-sampling at the lower end of the age distribution.[1] There would have to be pretests for all measures, including health measures. Only optional pretests are recommended for the health area, for reasons stated in Chapter 3.

The value-added design is certainly preferable to Design 6 and probably to Design 3. It seems preferable to Designs 4 or 5. It apparently yields less biased estimates of treatment effects than ANCOVA and requires no control groups. Thus, it is a less expensive design and can be used in sites that do not have enough eligible children to create a control group.

The idea has several known disadvantages, and since our experience with it is limited, undoubtedly some unknown disadvantages as well. We have had experience with constructing growth curves and time series, but we do not know how treatment estimates from random assignment and value-added designs compare. Of the disadvantages, any systematic variation between children as a function of age, excluding maturation, biases the growth curve. Since we propose to establish the growth curve cross-sectionally, one potential bias is what Hilton and Patrick (1970) call the cohort difference. The time lapse between pretest and posttest in the Head Start evaluation would not exceed seven months. Thus, children's posttest scores will be compared with pretest scores of a cohort seven months older. Certainly the two cohorts would have been born at different times of the year. It is not clear that season of birth is associated with systematically different children. Historical events might be associated with systematic differences if it can be demonstrated that (1) a widespread event occurred that could be expected to affect very young children, either directly or through their parents; and (2) the event occurred prior to the younger cohort's birth and after the older cohort's birth or the event could be expected to affect children differing only by a few months in age differently.

A possible source of systematic bias would be an interaction between age and selection. It is possible that children younger than the usual admission age are more capable than those of the usual age, and older children less capable. A child who is younger than the usual age may be admitted because he is able to "handle" himself; a child who is older may be admitted late because he lacks that ability.

---

[1] The discussion of stratifying the sample, below, notes the implications of an age basis for stratifying centers.

A third, less probable, source of bias associated with age is if there is a systematic difference in families associated with age of the child. For example, if there were a widespread rural-urban migration of young low-income adults, the evaluation could occur at the time their first children were entering Head Start. We know there are systematic differences between migrants and nonmigrants. If the older children in the centers that year were more often from nonmigrant families, there could be systematic differences between younger and older children in the catchment areas to which and from which the adults moved.

Another problem is variability around each point in the curve. For any given age, there is considerable variation in scores. This variation gives us less precise estimates of treatment effect for children in specific catchment areas, relative to untreated eligible children in that same area. As indicated earlier, there is considerable intersite variation in outcomes. Thus, for assessing the overall effects of Head Start, we feel more comfortable with aggregating site-specific effects than with comparing treatment children either (1) pooled across sites with control children pooled across sites, or (2) against growth curves based on pretest scores pooled across sites.

Another problem is the necessity of comparing posttest scores of children at the upper end of the age distribution with a projected, not an empirical, point on the curve.

Importance of Random Assignment. We strongly recommend Design 1 with volunteers randomly assigned to treatment and control groups, and within the treatment group, randomly assigned to classrooms.[1] Designs

---

[1] If children are normally assigned to classes on the basis of age, and the treatment is adjusted to the age group, random assignment to classes could distort the usual treatment situation in theoretically important ways. In this case it might be wise to allow the usual age allocation to classes if the age criterion is stated explicitly and assignment occurs on this basis only. The subsequent analysis of site data should then stratify by classroom.

2, 4, and 8 might be used in a limited situation described below. The reason for recommending Design 1 is that it allows credible site-specific estimates of effect (i.e., it requires random assignment of the total set of children to T and -T groups). Random assignment is less important in either of two circumstances: (1) when it is difficult to obtain random assignment and little or nothing is known about the subject of the evaluation, or (2) when the phenomenon to be evaluated falls within the scope conditions of a strong (consensually accepted) theory. In the first case, information from a quasi-experimental (nonrandom assignment) design may be preferable to no information at all. In the second case, there is a basis for predicting the results of the evaluation. To the extent that the predicted and observed results are consistent, they are credible.

Neither case entirely pertains to the Head Start evaluation. We have substantial information about cognitive effects of Head Start--e.g., from the Westinghouse-Ohio University and HSPV studies. The measures of cognitive effects also tap aspects of linguistic and perceptual-motor competence. We do not have much information about health and nutrition and socioemotional effects of Head Start. Unfortunately, differences between noncomparable treatment and control groups on either health and nutrition or socioemotional outcomes are plausibly explained by family variables associated with volunteering for Head Start. Enrolling the child in Head Start might indicate special concern or caring for the child. In this case it can be argued that the Head Start children would have been healthier regardless of Head Start: Parents who care enough about their children to enroll them in Head Start also care more about their children's health. In the socioemotional case, it can be argued that parents who show the initiative to enroll their children in Head Start are themselves socially more effective and serve as better models for their children than parents who do not enroll their children.

The other circumstance under which random assignment is less important, presence of strong theory, also does not pertain to the Head Start evaluation. With regard to the Follow-Through and HSPV evaluation, Marian Stearns (1974, pp. 15, 18) notes:

> An experiment, as everyone knows, requires something called randomization. Random assignment of treatments to subjects is done in hopes that all factors affecting outcome other than the treatments of interest have been equally distributed across experimental groups. An evaluator could proceed to collect data under circumstances where there was no randomization and could assume that valid inferences about causes be made, if he knew what all the other conditions were which affect the outcome and he knew how all of them work. There is no need for everything to be controlled or 'randomized out' in circumstances where the effects of all disturbing factors are known. But who can conceive of this situation existing in Follow Through?
>
> The poorer the theory or model, the more we must be able to make random assignment of treatments. If neither of these conditions is well met, then a planned variation evaluation should not be conducted.

Marshall Smith (1974, pp. 6, 8) states:

> Suppose an educational program is identified and tried out and a substantial change is observed in student outcomes. Before the results are attributed to the treatment we must be persuaded that the experiment was internally valid; i.e., that the results did not arise from something other than the treatment. Internal validity refers to the logic of the specific experimental situation--can we imagine plausible hypotheses other than the treatment which might causally explain the effect we would like to attribute to the treatment? Placed in this context we must reject an experiment which involves only one group--the group which receives a treatment. If we have only one group we will be able to imagine all sorts of rival hypotheses to explain differences between the pre- and post-treatment scores on an outcome measure. Other school influences, influences in the home, and biological maturation are all candidates. We are therefore forced to go to a multiple group experimental design, the simplest of which is a two group design.
>
> The internal validity of the two group design rests on the degree of assurance with which we can say the two groups differ only with respect to their receiving the particular treatment. We want to say that the difference in group outcome scores is unambiguously attributable to the fact that one group received the treatment (the treatment group) while the other group (the control group) did not. Another way of expressing this uses the term confounding. Systematic differences between the two groups other than their exposure to the treatment represent confounding influences--all confounding influences are rival explanations for the observed effects.

> Whatever process we use to remove confounding influences, use of probability statistics rests on the assumption that we have been successful. If we do not use randomization and instead rely on physical (matching) or statistical (covariate) procedures both our imaginations and our knowledge of the lack of adequate theory in the field make us realize that we must be less than successful.

Campbell and Erlebacher (1970, pp. 222-223) argue:

> Cause-effect interpretations of simple correlations gain credence when one is able to rule out the other available plausible rival hypotheses, and produce corroborating evidence.... The correlation of Head Start experience with subsequent ability scores...is swamped with plausible third variable explanations, both general environmental and genetic. This is exactly the problem acknowledged when one--as was done in the Westinghouse/Ohio study--uses matching and analysis of covariance to 'adjust' for possible alternate causes. These techniques, however, are demonstrably inadequate to the task, and have a systematic direction of bias.

In his analysis of the third-year HSPV data, Weisberg (1973, pp. 76-77) states:

> Undoubtedly the most serious design problem in this study is the lack of *randomization*. If a group of experimental units is divided randomly into two or more groups, then providing the groups are sufficiently large, there is only a small probability that they differ significantly on any given variable, measured or unmeasured. Of course, we can never be sure that the groups are equivalent with respect to *all* variables, but randomization is our best protection that there are no relevant group differences. If allocation to treatment groups is random we can be fairly confident that comparisons among group outcomes are unbiased even if no explicit account of pre-treatment variables is taken. We may still wish to use pre-treatment information to increase the precision of our comparisons, but with random allocation this information is more a luxury than a necessity.

The implication of prior experience with Head Start evaluations is that *if the proposed evaluation cannot yield unequivocal data, it is better not to conduct it.*

Implications of Random Assignment for Types of Centers. If Design 1 were used, Head Start centers could be divided into four categories:

1. Centers that serve all eligible children in the catchment area.
2. Centers that supply all demand, but do not serve all eligibles in the catchment area.
3. Centers that do not supply all demand, and are unwilling to allocate Head Start randomly among the demand.
4. Centers that do not supply all demand and are willing to allocate Head Start randomly among the demand.

The proportion of centers in each of these categories is unknown.

Category (1) does not allow a comparison group from the same catchment area. The only design solution for evaluating centers of this sort is Design 8.

On the bases of *current* local conditions, only category (4) centers are consistent with Design 1. Designs 4 and 8 are candidate solutions for category (2); Design 8, the preferred solution. Designs 2, 4, and 8 are the obvious ones for category (3), Designs 2 and 8 being preferred to Design 4. However, Design 1 is *theoretically* possible in categories (2), (3), and (4). Using this design requires increasing demand in category (2) and finding solutions to the problems that category (3) centers perceive. Recommendations for implementing random assignment are stated below. This discussion may touch on some of the problems for categories (2) and (3).

If centers from categories (2) and (3) cannot be shifted into category (4), we recommend these solutions in descending order of preference:

- If approximately 25 percent of the centers in the sample can be assessed with Design 1, use Design 8 to assess effects for these centers. This allows comparison of estimates from a value-added design with one from a random assignment design. Although Design 8 seems promising, it is important to evaluate it for bias.

- Use Design 4 if: (1) statistical linear models can be constructed to represent the expected sources of variation in outcomes for centers from categories (2) and (3); and (2) unbiased estimates of the parameters of the models can be obtained. If these conditions can be satisfied, Design 4 can then be used to collect data for centers in category (2), and Designs 2 or 4 for centers in category (3). To the extent that results for these centers are consistent with results predicted by the models, they are *considerably* more interpretable and credible than results based on Designs 2 or 4 alone. This strategy requires a special methodological study to specify models and estimate their parameters. The study should be embedded in the national evaluation and the parameters estimated from data collected by means of Designs 1 or 2.
- Do not evaluate centers in categories (2) and (3). This means dropping these centers from the sampling list of centers. It also means that results of the evaluation can be generalized only to category (4) centers. To the extent that significant numbers of centers fall into categories (2) and (3), the evaluation sample is no longer even approximately representative. The decision to proceed with such an evaluation depends on its point. If OCD is primarily concerned with knowing the effectiveness properties of the total Head Start system, sample representativeness is important. In this situation the evaluation should probably not be conducted. If OCD is primarily concerned with estimating the effect of Head Start as a *concept*, the evaluation should proceed. Under these circumstances what is of primary importance is an unequivocal estimate of effects of Head Start instances, regardless of their representativeness.

Recommendations for Conducting Random Assignment. The basic procedure is to assign randomly from a list of volunteers for Head Start, where the assignment is not only random *between* T and -T groups, but between classrooms *within* the T group.

Head Start personnel may know from previous experience that they will not obtain a sufficient number of volunteers to create T and -T groups of recommended size. If there is an excess of eligibles over volunteers, the nonvolunteer eligibles have not signed up for any of the following reasons: (a) the child's family is not interested in Head Start; (b) the child's family does not know something crucial about the program--e.g., is existence, their eligibility, registration dates; (c) it is difficult for the family to register the child--e.g., transportation, baby-sitting, or work schedule problems. The center should be helped to run its recruiting program so that it alleviates reasons (b) and (c). These are *artifactual* reasons for not volunteering. Special recruiting efforts should not be set up to affect reason (a). The evaluation is restricted to children whose families volunteer them for the program. The important difference between volunteers and nonvolunteers is family interest. To the extent that children from uninterested families are obtained for the evaluation sample, that sample can be expected to be different from the ordinary group of Head Start children.

Head Start centers generally do not use random assignment to treat excess demand. Thus, asking centers and Head Start parents to use this procedure may be asking them to use a different procedure for distributing a scarce resource. Random assignment creates at least three problems: it increases the administration work of Head Start personnel, interferes with local autonomy, and changes the equity ground rules for distributing a scarce resource. These are legitimate problems for Head Start personnel and parents. Thus, it is imperative that random assignment be conducted so as to minimize and compensate for these problems. At the same time, *if* constituents to the evaluation, including Head Start personnel and parents, think the idea of the evaluation is sound, it is important that the data be as valid as possible. Invalid data are no help to any constituent: We either learn nothing, or we learn the wrong things. This is not fair to the children, their parents, personnel dedicated to the program, OCD, or taxpayers and their representatives.

We suggest the following ways to minimize or compensate for the problems that random assignment poses for Head Start personnel and parents.

- Members of the evaluation team--e.g., the site coordinator--should assume as much of the administrative work of random assignment as possible. Head Start personnel can be offered at least two compensations for their additional work. One is to give each center two or three choices of measures that are administered specifically for their information, not for the national evaluation. The other is to provide them with a special analysis and interpretation of the basic battery data for their site alone. No public report on the evaluation can contain data on an identifiable site or child. However, if personnel at a specific Head Start center want the results for their center alone, they should have access to those data. They also should assume the obligation to handle its dissemination as they wish.
- Head Start is a program that mandates local options and initiatives. To ask Head Start personnel and parents to change selection procedures is to interfere with that autonomy. This can only be done with local concurrence. They have a right to know why they are being asked to use a different procedure and to have veto power over its institution. As discussed in Chapter 9, perhaps a neighborhood meeting for Head Start personnel and parents could be scheduled to discuss the evaluation in general and random assignment in particular. If Head Start personnel and parents decide to use random assignment, one way they can control its use is to set up a lottery, with parents drawing names for the T and -T groups and for classroom_1 and classroom_2 in the T group.

o There are several equity issues. One is the basis for distributing Head Start services. Two frequent ways of handling excess Head Start demand are first-come/first-served, and need. Thus, if Head Start personnel and parents accept any grounds for refusing Head Start, they are apt to accept arrival time or differential need. Random assignment ignores both bases.

We argue that neither basis is necessarily equitable. The first-come/first-served basis favors those most knowledgeable and those for whom registering the child is easiest. The need basis has other difficulties. The validity of need indicators is unknown. Almost inevitably, less needy children are admitted and more needy ones turned away. Random assignment is not clearly less equitable than either selection basis; it is undoubtedly more equitable in some cases.

Other equity issues involve fairness to later participants in Head Start and current participants in the evaluation. If Head Start personnel and parents feel that the evaluation can improve services to their children, it is only fair to subsequent Head Start participants to collect valid data now. Similarly, if Head Start personnel and parents agree to participate in the evaluation, it is only fair that their participation and that of their children be worthwhile--that the evaluation yield valid data.

In sum, random assignment may produce more short-term and long-term equity. However, Head Start personnel and parents should be compensated for entertaining alternative ground rules. If Head Start personnel use need as a basis for selecting children, we strongly recommend that they be given the opportunity to evaluate the validity of their indicators. One way to do this is as follows. Head Start personnel carry out their usual selection procedure to the point of compiling a list of children whom they would ordinarily admit. If they feel that what they are trying to predict is reflected in the dependent variables of the evaluation, the T and -T children's pretest scores allow them to check the accuracy of their prediction. If there is a high correlation between the list of children they ordinarily would have selected and low pretest scores, the selection criteria would seem to be valid.

Parents can also be compensated. They have the right to an explanation of results for their children, including uncertainties in their interpretation. A second compensation might be to provide followup services to the health and nutrition posttest. Since -T children receive the health and nutrition battery, one benefit for parents of these children is diagnostic information on the child. However, the evaluation staff, given resources, could also arrange relevant medical followup for each child. This benefit is possible if the evaluation avoids a longitudinal estimate of health and nutrition effects. If OCD wants to run a longitudinal evaluation, both T and -T children are contaminated by followup and by diagnostic and health inputs.

STRATIFICATION DIMENSIONS FOR TREATMENTS AND CHILDREN

There are statistical and policy reasons for stratifying both treatments and eligible children. This section discusses the reasons and criteria for stratifying, choice of treatment dimensions, and choice of children dimensions.

Rationale and Criteria

The statistical reason for stratifying is straightforward: The sampling error of the estimate of the population mean on a measure derives entirely from within-stratum variation. If we understand the major sources of variation in the population for that measure, we can sort the population into categories (strata) that maximize between-strata and minimize within-stratum variation. This decreases the size of the error term and consequently increases the power of statistical tests (i.e., increases the probability of rejecting the null hypotheses when the alternative is true). If we do not understand the major sources of variation in the population for that measure, an unstratified design is preferable. In this situation, there is no gain from stratification (i.e., the within-stratum error terms will not be reduced because we do not know how to detect homogeneous elements). Stratification that does not create homogeneous groups has two costs: the time and money required to classify elements of the sample according to stratifying dimensions; and--*if* the data are analyzed as stratified data--less powerful statistical tests.[1] A stratified or "block" design implies fewer degrees of freedom for the error term. Degrees of freedom reflect sample size. As indicated below, sample size is positively related to statistical power. Tests on stratified samples are based on the stratum sample size. Since the sample size for a stratum is necessarily smaller than the size of the unstratified sample, statistical power is lower for stratified samples unless there is a reduction in the experimental error term.

The policy reason for stratification is also straightforward. An estimate of the effects of Head Start on eligible children in general provides little useful information for constituents of a national evaluation. This information is useful for constituents concerned with alternative allocations of resources--e.g., the Office of Child Development, Office of Management and Budget, the Congress. However, it is not sufficient information even for these groups. The effects of Head Start on eligible children in general may be quite different from effects on particular groups of children. It is improbable that a program as diverse as Head Start has uniform effects for a group as diverse as low-income children. If evaluations cannot discriminate effects for different groups in the target population, they cannot give policymakers an adequate information basis for making other than a binary choice: Keep the chaff, as well as the wheat, or throw out the wheat and the chaff.

---

[1] There is nothing necessary or irreversible about this consequence. Although the sampling frame may be a stratified frame, the data do not have to be analyzed as stratified data. If they are so analyzed, they can always be reanalyzed as unstratified data.

Other constituencies--e.g., Head Start parents and program personnel--are not concerned with children and programs in general. They are responsible for particular children and particular programs. They need to know: What are the effects of our programs on our children? where "our" is defined more specifically than Head Start in general or eligible children in general.

In sum, to provide useful information for all constituents of a national evaluation, the evaluation must be able to yield statements about effects of types of treatments on types of children. It is possible to make these statements only if the samples of relevant treatment-children combinations are large enough to yield reliable estimates of effect. Stratifying the treatment and child populations allows us to *ensure adequate cell sizes efficiently*.

Statistical and policy reasons for stratifying produce different criteria for selecting dimensions. On statistical grounds it is necessary to select dimensions of treatments and children that correlate with variations in effect; on policy grounds, it is essential to select dimensions that relate to groups organized to *act on* information from the evaluation. These different criteria may imply different dimensions. The statistical costs of inappropriate dimensions have been stated. The policy costs are that information from the evaluation is less apt to be useful (i.e., less able to be translated into action).

Choice of Treatment Dimensions

Unfortunately, there are two serious problems with stratifying treatments. First, we have no theoretical basis for knowing which classroom variations make a difference. We consider three aspects of this problem:

- Head Start Planned Variation set up treatment variations and evaluated their effects. Thus, there are estimates of the effects of specified variations in treatment. However, there are serious methodological problems with those estimates. That experiment also involves only 210 classrooms. These variations were never systematically diffused throughout the population of Head Start classrooms. Thus, there is no reason to expect much resemblance between what occurs in most classrooms and what occurred in a small set of experimental classrooms.

- There is a general literature that investigates the relationship between variations in educational process and variations in effects on children. However, no compelling theory of classroom process has emerged from this literature.
- Chapter 6 of this report indicates process variables that might prove promising. If the pilot test of the evaluation (see Chapter 9) indicates that these variables strongly predict variations in Head Start effects, they can be used as stratifying dimensions for treatment.

The second problem with stratifying treatments is relevant only if the first problem is solved. This is the problem of determining which classrooms fall into different treatment categories. If the first problem is solved, it will probably be because the process variables of Chapter 6 successfully discriminate programs. The information for categorizing the population centers on these dimensions is not now available at any administrative level of the Head Start program. This categorization problem is traditionally handled by *double sampling*. In other words, depending on the desired final sample size, centers are oversampled by some order of magnitude. Only data on process variables are collected in the oversample. Centers are then categorized into treatment categories. The final sample is selected from these categories.

Double sampling is expensive. It is worth considering only if the pilot test of the process variables indicates that these variables correlate with variations in effects.

In sum, *for stratifying treatments, we recommend analysis of the pilot data for a relationship between process and effect variations.* If there is a systematic relationship and double-sampling is economically feasible, an unstratified sample of centers n times the size of the desired final sample should be drawn,[1] the centers categorized on relevant process variables and the final sample drawn from these categories. If there is no systematic relationship, treatments should be stratified into two levels: Head Start and non-Head Start (i.e., treatment and control). In this case, the center population is unstratified by treatment. In Chapter 8, we recommend a data analysis strategy for dealing with assumed but unnamed variation in treatments.

Choice of Children Dimensions

The objective of stratifying children is to create groups of children who are socially more identifiable and statistically more homogeneous than the total set of Head Start children.

As indicated, stratification is useful statistically only if selected variations in children correlate with variations in Head Start effects. Selecting such dimensions is somewhat complicated by multiple dependent variables. In other words, we are not interested in dimensions that create homogeneous groups of children for only one dependent variable. Statistically, it is worth stratifying the sample only if the dimensions create fairly homogeneous groups of children for the *set* of dependent variables.

One way to select dimensions is to analyze data from the pilot run of the evaluation by a procedure such as discriminant analysis. Is there a small number of distinct collections of scores on the dependent variables? If so, can we recognize members of these clusters in advance of the treatment? In other words, can we sort children into categories associated with different clusters of scores on the dependent variables? The nature of the score clusters may provide clues about differences between children. However, the task of "naming the cluster" or "naming the factor" is an *ex post facto* and consequently precarious exercise. It is possible that children's scores on pretests of the dependent variables fall into distinct clusters and that pretest and posttest clusters are highly associated. If so, children's pretest scores can be used to stratify children, at least for purposes of the statistical analysis of posttest results.

---

[1] The value of n is determined by the number of treatments, the observed frequency of each in the pilot sample, the observed frequency relative to types of children, and the desired cell size for each treatment-child combination to be evaluated. In the unlikely case that treatment types are equally represented in the pilot sample and each type distributes evenly across types of children, n = 1.

From the policy perspective, stratifying on clusters of pretest scores is useful for program personnel who wish to adjust the treatment for children who score in a particular way. However, it has much less national policy appeal. Groups of children may be differentiated by how they score on, say, the CIRCUS test What Words Mean. But who are these children? Identification of and with individuals in the social group occurs in terms of the parameters of our experience: where we live, ethnicity, economic status, age, sex, etc. Variations on these dimensions produce fundamental variations in interests between groups. Pressure groups are tied to these different interests, and social policies are designed around them. If constituents are to act on information yielded by the evaluation, that information has to be related to these dimensions of social experience.

On the basis of the policy criterion for stratifying the sample, the dimensions should be the basic demographic distinctions among Head Start-eligible children. If selected demographic variables demarcate fundamentally different experiences, these differences in experience may be systematically associated with different responses to a Head Start treatment. If so, the demographic variables sort children into fairly homogeneous groups with regard to the dependent variables. In that case, the demographic variables satisfy policy and statistical criteria for choosing stratifying dimensions. However, this association cannot be assumed. It should be tested on the pilot test data. If clustering on dependent variables cuts across demographic variables, then for *drawing the sample of children*, the population should be stratified on the basis of demographic variables. For *analyzing the data*, the within-site sample might be stratified according to dimensions associated with dependent variable clusters; or if these dimensions are unidentifiable, the sample should be unstratified. Each Head Start center is connected with treatment instances (classroom) and demographically a fairly homogeneous catchment area. To observe the effects of treatments on children with different demographic properties, it makes sense to sample from the population of Head Start centers, stratified according to demographic properties of their catchment areas.

We suggest the following stratifying dimensions. In evaluating their adequacy, the reader should keep in mind that we want to be able to estimate the effects of Head Start separately for each stratum, as well as across strata. In a site by site analysis strategy sample size for a stratum is determined by the number of sites whose catchment areas have the demographic properties of that stratum. Although attrition should be lower for sites than for children, there should be ten sites per stratum (see discussion of sample size below). This limits the number of strata that can be created. For example, if we assume a minimum sample size of ten per stratum, ten strata require a total sample size of 100 centers.

We recommend three stratification dimensions for Head Start-eligible children and discuss each of these in turn: (1) cultural nature of the group; (2) location of the group on a metropolitan/sparseness dimension; and (3) regional location of the group.

Cultural Dimension. We recommend the following cultural distinctions: Native American, Chicano, Puerto Rican, Black, and White. These groups are selected on the basis of four criteria:

1. Recent or continued political disenfranchisement.
2. Fundamental differences in parent culture patterns.
3. Proportion of the group served by Head Start.
4. Self-defined distinctness.

Criterion (1) is included because if a child's social group has been systematically disenfranchised, that child has limited access to mainstream opportunities and self-limits his acceptance of those opportunities that are available. Head Start is seen by many as a way to equalize opportunities among children. It is then particularly important to see the effects of Head Start on children whose opportunities are seriously limited.

Criterion (2) is included as a proxy for variations between children in the strengths and problems they bring to Head Start. Criterion (3) is an efficiency criterion. Funds for evaluation buy information; they are also limited. Limited funds are best spent to buy information for the largest groups of Head Start-eligibles that are maximally distinct from one another. Criterion (4) is a political criterion. On cultural and social grounds, differences between subgroups of one category may be no greater than those between subgroups of other categories. However, they may perceive themselves and wish to be treated as distinct.

Native American children constitute only about 2.5 percent of the total number of children served by Head Start. However, the Native American group is distinguished from the others on the basis of criteria (1) and (2). The Black group is distinguished on the basis of criteria (1), (2), and (3). The White group is distinguished for reasons (2) and (3). The Spanish-speaking culture is distinguished from the White group on the basis of criteria (1) and (2). The reasoning is as follows. The White group includes ethnically different groups. However, the Spanish-speaking groups are culturally more different from and politically less enfranchised than White groups: there is more recent immigration to the United States in the case of Puerto Ricans, and geographical proximity to parent cultures in Puerto Rico and Mexico.

Puerto Ricans and Chicanos are distinguished from each other on the basis of criterion (4). It is true that Chicanos and Puerto Ricans come from different branches of Spanish culture in the New World, and that their experiences in the United States are different. Chicanos resided in the Southwest when those lands were Spanish territories. They remained distinct from Anglo culture, partly because their populations were sizable and contiguous to the parent culture. They also remained distinct during a period in American social history when assimilation, not pluralism, was valued.

The Puerto Ricans recently migrated to the United States (beginning in the late 1940s). They were citizens, not immigrants, as recent Chicano arrivals are; their experience in the United States has been only urban. There are Puerto Ricans of dual Black and Spanish origin, which confounds their experience with Whites. The effects of cultural difference and political disenfranchisement have probably been different for the two groups. First, Puerto Ricans arrived at a time when pluralism was beginning to be entertained as a positive social possibility.

Thus, cultural difference in Puerto Ricans was probably less negatively evaluated than it was for Chicanos. Second, although Chicanos and Puerto Ricans are both in the process of obtaining political power, the Puerto Ricans are doing so after two or three generations; the Chicanos, after many.

Nevertheless, child-rearing practices are the same for the two groups. Other differences between them are probably not larger than those between subgroups in other categories not differentiated here. We distinguish the two groups as different, but that is also true of, for example, the Irish and the Italians. The reason to differentiate Chicanos and Puerto Ricans is political. At this point in their social history the two groups consider themselves and wish to be treated as distinct.

Certain groups are not distinguished for the same four reasons. For example, the Cubans and the Chinese are few in total number and still fewer in number of Head Start-eligible families (criterion 3). The diverse cultural groups within the White group are not differentiated for reasons (1), (2), and (3). The parent cultures of these groups are less different from each other than from, for example, any one culture of the Native American tribes. The immigrations of these groups are not recent; in most cases the children of Head Start age are at least third and fourth generation. Thus, there has been time for commonality to develop between the more recent immigrations of Irish, Italian, Swedish, German, Polish, Jewish, and Hungarian groups, and the earlier English, Scottish, Dutch, and French immigrations. There has also been time for political power to be distributed among these groups. Finally, any one of these groups represent a very small proportion of the group served by Head Start.

Metropolitan/Sparseness and Regional Dimensions. As an indicator of urban influence, the traditional urban/rural distinction does not reflect the recent shift in urban influence patterns in the United States. *Rural Development Goals: First Annual Report of the Secretary of Agriculture to the Congress* (January 18, 1974) presents an alternative. The report lists seven county categories:

County Type	Definition
Metropolitan (Standard Metropolitan Statistical Area, or SMSA)	The technical federal definition of SMSA appears in several sources, including the U.S. Bureau of the Census *City and County Data Book*. Briefly, it refers to all counties that (1) contain a city of ≥ 50,000 residents or "twin cities" with a combined urban population of ≥ 50,000; or (2) are contiguous to a county with such a city and are socially and economically integrated with that city.
Nonmetropolitan	All counties not included in the above.
Urbanized Adjacent	Counties contiguous to SMSAs and containing 20,000 to 49,999 urban residents. An "urban resident" is a person who resides in an urbanized area, as defined by the U.S. Bureau of the Census.
Urbanized Not Adjacent	Counties not contiguous to SMSAs and containing 20,000 to 49,999 urban residents.
Less Urbanized Adjacent	Counties contiguous to SMSAs and containing 2,500 to 19,999 urban residents.
Sparse Adjacent	Counties contiguous to SMSAs and containing less than 2,500 urban residents.
Sparse Not Adjacent	Counties not contiguous to SMSAs and having less than 2,500 urban residents.

Table 7-2 shows distribution of total population according to county categories, distribution of total low-income population according to regional and county categories, and birth rate according to county categories.

Useful *density* distinctions vary with the cultural group. Black and Chicano Head Start-eligible families reside primarily in the central cities of Standard Metropolitan Statistical Areas (SMSAs) or in nonmetropolitan areas. Few reside in the metropolitan ring. Puerto Rican Head Start-eligible families reside almost entirely in central cities.

Table 7-2

SELECTED CHARACTERISTICS OF COUNTY GROUPS

Characteristic	Total All Counties	Metro-politan	Nonmetropolitan					
			Urbanized		Less Urbanized		Sparse	
			Adjacent	Not Adjacent	Adjacent	Not Adjacent	Adjacent	Not Adjacent
1. 1970 population								
Thousands	203,213	147,996	13,967	7,644	13,307	13,598	2,325	4,515
Percent of Total	100.0	72.8	6.9	3.8	6.6	6.7	1.1	2.2
2. No. of counties	3,097	612	191	137	564	721	246	626
3. Incidence of poverty, 1969:								
Thousands	27,840	16,724	2,053	1,368	3,215	3,028	621	1,237
Percent of Total	13.7	11.3	14.7	17.9	21.0	22.6	26.7	27.4
Northeast	10.1	9.7	11.6	13.2	12.6	14.0	14.3	13.4
North Central	10.8	9.2	10.4	13.2	13.2	16.5	17.5	21.3
South	20.3	15.7	21.1	23.5	28.3	30.1	32.1	35.3
West	11.7	10.7	15.9	14.4	16.5	15.6	14.6	21.8
4. Children born per 1,000 married women 35-44 years old	3,132	3,040	3,259	3,293	3,431	3,476	3,572	3,654

SOURCE: *Rural Development Goals: First Annual Report of the Secretary of Agriculture to the Congress*, January 18, 1974, Tables 1, 35, 39, 46.

The federally funded Head Start program for Native American children is restricted to Native Americans on reservations. The reservation unit does not fit into county categories. For reasons discussed below, the minimum regional distinctions for Native American groups are too numerous to use as a basis for stratifying Native American groups into more homogeneous sets. Two dimensions that seem crucial are the tribe's isolation and traditionalism. The first dimension seems important because geographic isolation compounds the political disenfranchisement of Native American groups. Under the best of circumstances they have limited opportunities. Geographical isolation would reduce those that might be available if they were proximate to larger settlements. The traditionalism concept seems important as an indicator of degree of cultural difference among tribes. The effect of Head Start might be expected to differ, depending on whether the child comes from a group that is closer or farther from the White culture. One dimension that might reflect the isolation and traditionalism concepts is frequency of a metropolitan contact. There are two problems with this dimension. One is definitional: How is "frequent" distinguished from "infrequent?" An appropriate definition should be worked out with knowledgeable Native American groups. The second problem is a validity question. Is variation in amount of contact with a metropolitan area highly correlated with tribal traditionalism? For example, some tribes do not have much contact with SMSAs, but run very successful resort services. At least economically these groups are much closer to White culture. The recommendation is to: (1) stratify Native American groups into frequent and infrequent contact with SMSA, and (2) ascertain the fruitfulness of this distinction through consultation with knowledgeable Native American groups on reservations.

As indicated below, aside from Appalachia and the Ozark region, urbanization of population seems to affect White culture experiences more than region. We recommend two analytic categories: central city and nonmetropolitan.

Useful *regional* distinctions also differ according to particular racial or ethnic groups. Although the difference between South and non-South is important for Blacks, the regional distinction duplicates central city and non-metropolitan distinctions. The central city experience of Blacks is fairly uniform across regions, South and non-South, and the majority of nonmetropolitan Blacks reside in the South.

The vast majority of Puerto Rican Head Start-eligibles reside in New York City and New Jersey. Thus, regional distinctions are not relevant for this group. Chicanos reside primarily in the Southwest, defined as Texas, New Mexico, Arizona, Nevada, Utah, Colorado, and California. There are sizable Chicano settlements in Milwaukee and Chicago. The recent arrivals, who are apt to be Head Start-eligible children, are similar to Chicanos in the Southwest. If there are insufficient central city sites in the Southwest, centers in Chicago and Milwaukee might be included in the population for this stratum. In that case, the stratum should be defined as "central city, Southwest, Chicago, and Milwaukee." In general, though, Chicanos who have moved out of the Southwest tend to move into the White culture.

Making a limited number of regional distinctions for Native American groups on reservations becomes a largely arbitrary exercise. A number of regionally based political coalitions correspond to original cultural commonalities. To the extent that they reflect common interests and problems, they could be useful bases for stratification. The groups are: Coalition of Eastern Native Americans, United Southeastern Tribes, Great Lakes Inter-Tribal Council, United Sioux Tribes of South Dakota, United Sioux Tribes of North Dakota, Northwest Affiliated Tribes, Inter-Tribal Council of California, Inter-Tribal Council of Nevada, All-Indian Pueblo Council, Arizona Inter-Tribal Council, United Tribes of Western Oklahoma and Kansas, Five Civilized Tribes of Eastern Oklahoma, and the Alaska Federation of Natives. There are two problems here. First, Head Start serves only a total of 10,000 Native American children. We may want approximately 10 sites per stratum and 30 Head Start children per site. If the 10,000 children are evenly distributed among the 13 strata, the sampling fraction is ≈ 0.4. Since field realities are apt to restrict the population of Head Start centers/catchment areas from which we can sample, it may be impossible to obtain adequate cell sizes for this number of strata. Second, estimating effects for 13 strata requires a disproportionately large number of funds for a disproportionately small number of Head Start children.

There are many notable differences within the White culture. For example, Georgia, Maine, and Minnesota farm families seem different culturally. However, the basic child-rearing patterns do not diverge widely. Appalachia and the Ozarks, however, stand out from all other regions. Since their extreme isolation is associated with a distinctive life style, experiences for Appalachian and Ozark Whites are suggested as the important distinction in the White group.

The following list stratifies the five cultural groups according to metropolitan/sparseness and regional criteria. It represents an exhaustive classification for Head Start centers/catchment areas. We recommend that effects be separately estimated for the categories on the list with an asterisk (*). In other words, these categories are sample categories.

Stratification of Cultural Groups[1]

1. Black
    *a. Central city, defined by the U.S. Census Bureau as any county containing a concentration of ≥ 50,000 residents
    *b. Nonmetropolitan
    c. Metropolitan, noncentral city

*2. Puerto Rican

3. Chicano
    *a. Central city in the Southwest, where "Southwest" is defined as Texas, New Mexico, Arizona, Nevada, Colorado, Utah, California
    *b. Nonmetropolitan in the Southwest
    c. Metropolitan, noncentral city in the Southwest and all other

4. Native American
    *a. Frequent contact with an SMSA
    *b. Infrequent contact with an SMSA

---

[1] A catchment area can belong to only one metropolitan/nonmetropolitan category and to only one regional category. Ethnically, the catchment area may not be homogeneous and thus may fall into two or more cultural categories. A decision rule should be established for categorizing the sites as a Black, Chicano, Puerto Rican, White, or Native American site. A possible rule is that if the catchment area is ≥ 75 percent one ethnic group, it is classified in a stratum for that group. If there is no clear preponderance of an ethnic group, the site can be eliminated from the sample or a residual category of "mixed cultural group" established. If ethnically mixed sites are eliminated from the sample, this rule becomes another restriction on the sampling list.

5. White culture
    - *a. Isolated Appalachia and Ozarks, defined as Kentucky State Economic Areas 5, 8, 9; Virginia SEA 1; West Virginia SEA 2, 4; Missouri SEA 7, 8; and Arkansas SEA 9.[1]
    - b. All other
        - *(i) Central City
        - *(ii) Nonmetropolitan
        - (iii) Metropolitan, noncentral city.

SAMPLE SIZE

Sample Size Decisions

Two sampling decisions have to be made for a national evaluation of Head Start: number of children per site and number of sites. These numbers are a function of the effects to be estimated and the data analysis strategy. The design is structured to estimate (1) effects of Head Start across treatment instances and children (*system* effects), and (2) effects of Head Start for each of several demographically identifiable groups in the eligible population (*stratum* effects).

The implications of these estimates for sample sizes for sites and children depend on the data analysis strategy. As indicated in the Introduction and discussed in more detail in Chapter 8, we do not expect pooling across sites to be either statistically advisable[2] or substantively interpretable. The recommended analysis strategy is (1) analysis of treatment effects for each individual site, and (2) aggregation of effects for all sites for question 1 and for stratum-specific sites for question 2. The implications of this strategy for sample size decisions are: (1) the number of children per site have to be sufficient to allow the site to stand as a separate experiment; and (2) the number of sites per stratum have to be sufficient to yield a stable estimate of proportion of sites successful, or mean effect, or whatever.

---

[1] This definition of Appalachia and Ozark was suggested by Mr. Calvin L. Beale, U.S. Department of Agriculture.

[2] Although pooling increases the degrees of freedom for the error term, it also probably markedly increases its size.

Criteria for Sample Size Decisions

The criteria for selecting sample sizes are to obtain *accurate* and *efficient* estimates of system and stratum effects, given a site-by-site analysis strategy. *Priorities* among the constituents to the evaluation become relevant to sample size decisions to the extent that constraints do not let us estimate all effects with the desired precision.

Accuracy of an estimate is related to the variability of T and -T and to the variability of the population for any particular value of T and -T variables. If a population is completely homogeneous on a variable, a sample of one unit yields information as accurate as a sample of 10 or 100 units. The greater the variability in a population, the larger the sample needed to estimate properties of that population accurately. "Efficient" is defined here as the minimum number of sampling units to ensure a designated probability of accuracy. Priorities can be substantive or methodological (i.e., content or quality of information). Constraints can be budget, manageability of field operations, number of classrooms at each center, etc.

The remainder of this section discusses the relationship between sample sizes and the precision of estimates and the possibility of optimum allocation. The chapter concludes with recommendations for sample sizes.

Relationship between Sample Size and Precision. Sample size affects accuracy of treatment estimates through the power of a statistical test of a research hypothesis.[1] Let us assume that for the question of general Head Start effect the null hypothesis is $T = -T$; the alternative, $T \neq -T$. The power of the test of this hypothesis is then the probability of rejecting the hypothesis that $T = -T$, when in fact $T \neq -T$. This probability is a function of six variables: (1) magnitude of the difference between T and -T means on dependent variable_i, (2) sample size, (3) number of treatment levels, (4) population error variance (e.g., variability of the population on dependent variable_i), (5) probability of rejecting the null hypothesis when it is true, and (6) extent to which stratification, pretest, and other background information about children can be used to reduce the variability of the population on the dependent variables.

---

[1] The power of a test is defined as the probability of rejecting the null hypothesis when the alternative is true.

Tables 7-3 and 7-4 are presented to give some sense of how these variables operate to affect power. Table 7-3 shows how power of a test varies with variation in sample size per cell for several CIRCUS measures: CIRCUS instrument No. 1 (What Words Mean), No. 3 (Look Alikes), No. 4 (Copy What You See), No. 13 (Think It Through), No. 16a (Behavior Inventory: Enjoyment), and No. 16b (Behavior Inventory: Inattentiveness). For CIRCUS No. 2 (How Much and How Many), Table 7-4 shows how power of a test varies with variation in magnitude of the difference between T and -T on dependent variable_i, population error variance, and sample size.

Both tables hold two variables constant: number of treatment levels, and probability of rejecting the null hypothesis when it is true. Treatment levels are held constant at two, defined as T (program condition) and -T (control condition). Probability of rejecting the null when it is true is held constant at 0.05.[1]

In Tables 7-3 and 7-4, statistical power is calculated for sample sizes from 10 to 30 per cell in increments of 5 and for 60, 100, 140, and 200 per cell. At the *site* level we can expect approximately 15 children per classroom before attrition and, although some centers have more classrooms, no more than two classrooms. Thus, we expect an upper bound of 30 children per site in the T condition. Assuming an equal number of control children, we expect an upper bound of 30 children per site in the -T condition. We can assume attrition of these two groups. The calculations for 10-30 per cell are intended to reveal the feasibility of a site-by-site analysis under different attrition assumptions. The calculations for 60-200 per cell are included to show the effect on power of pooling data across sites for the analysis. The variations in number are determined by variations in assumed attrition and number of sites for which data are pooled. As indicated, we do not expect to pool data across sites. However, the pilot test of the evaluation may

---

[1] Chapter 8 discusses bases for selecting alpha levels in more detail. We use a level of 0.05 here because it is a conventionally acceptable level.

Table 7-3

STATISTICAL POWER[a] FOR CIRCUS TESTS NOS. 1, 3, 4, 13, 16a, AND 16b

Sample Size Per Cell	Test Number[b]					
	1	3	4	13	16a	16b
10	≃ 0.22	≃ 0.22	< 0.10	≃ 0.18	≃ 0.22	< 0.10
15	≃ 0.30	≃ 0.30	≃ 0.10	≃ 0.26	≃ 0.30	< 0.10
20	≃ 0.40	≃ 0.40	≃ 0.11	≃ 0.32	≃ 0.38	≃ 0.12
25	≃ 0.49	≃ 0.48	≃ 0.14	≃ 0.39	≃ 0.47	≃ 0.14
30	≃ 0.55	≃ 0.56	≃ 0.16	≃ 0.45	≃ 0.54	≃ 0.15
60	≃ 0.85	≃ 0.85	≃ 0.25	≃ 0.73	≃ 0.84	≃ 0.25
100	≃ 0.97	≃ 0.97	≃ 0.39	≃ 0.91	≃ 0.97	≃ 0.37
140	0.99	0.99	≃ 0.51	≃ 0.98	0.99	≃ 0.48
200	0.99	0.99	≃ 0.65	0.99	0.99	≃ 0.62

[a] Tang's formula (1938) is used to calculate statistical power. The parameter ϕ is defined as:

$$\phi = \sqrt{\frac{\sum_{1}^{k} (\mu_j - \mu)^2/k}{\sigma_\varepsilon/\sqrt{n}}}$$

where μ = total mean of treatment populations,

μ_j = mean for a specific treatment level,

k = number of treatment levels,

σ_ε = square root of the population error variance, and

n = sample size per cell.

$k = 2$.

[b] μ_1 is defined as the treatment group (T); μ_2, as the control group (-T). For test No. 1, $\mu_1 = 27.8$, $\mu_2 = 26.6$, $\sigma_\varepsilon = 5.95$; for No. 3, $\mu_1 = 18.8$, $\mu_2 = 16.1$, $\sigma_\varepsilon = 4.95$; No. 4, $\mu_1 = 31.80$, $\mu_2 = 30.1$, $\sigma_\varepsilon = 7.15$; No. 13, $\mu_1 = 21.30$, $\mu_2 = 18.6$, $\sigma_\varepsilon = 5.7$; No. 16a, $\mu_1 = 8.6$, $\mu_2 = 7.3$, $\sigma_\varepsilon = 2.4$; and No. 16b, $\mu_1 = 15.5$, $\mu_2 = 14.5$, $\sigma_\varepsilon = 4.35$.

Table 7-4

STATISTICAL POWER FOR CIRCUS TEST NO. 2:
HOW MUCH AND HOW MANY[a]

Sample Size Per Cell	Unreduced Variance[b]		Reduced Variance[c]	
	1-Step Effect[d]	2-Step Effect[e]	1-Step Effect	2-Step Effect
10	≃ 0.17	≃ 0.36	≃ 0.30	≃ 0.67
15	≃ 0.23	≃ 0.52	≃ 0.43	≃ 0.84
20	≃ 0.29	≃ 0.65	≃ 0.54	≃ 0.93
25	≃ 0.35	≃ 0.75	≃ 0.65	≃ 0.98
30	≃ 0.40	≃ 0.83	≃ 0.73	0.99
60	≃ 0.69	≃ 0.98	≃ 0.95	0.99
100	≃ 0.89	0.99	0.99	0.99
140	≃ 0.96	0.99	0.99	0.99
200	0.99	0.99	0.99	0.99

[a] For middle versus low SES: $\mu = 26.04$, $\mu_j = 24.49$, $\sigma_\varepsilon = 6.91$; for high versus low SES: $\mu = 27.1$, $\mu_j = 24.49$, $\sigma_\varepsilon = 6.89$.

[b] Unreduced variance = mean of standard deviations for all three SES groups.

[c] Reduced variance = 2/3 mean of standard deviations for all three SES groups.

[d] One-Step effect = middle-SES mean - low-SES mean.

[e] Two-Step effect = high-SES mean - low-SES mean.

reveal that (1) sites cluster with regard to community effects, child background characteristics, and variants of T and -T; and (2) these clusters are identifiable in advance of the full-scale evaluation.

In order to obtain estimates for the other two variables, magnitude of difference between T and -T means and population error variance, we used the normed data for several CIRCUS measures that we propose to use in the national evaluation. These measures were normed on a sample of nursery school children, selected randomly by a cluster sampling procedure. The means and standard deviations were calculated by various properties of the children. One of these was SES of the child's family.

We assume that Head Start-eligibles fall into the low-SES group, as defined by ETS. In order to get an estimate of Head Start effect, we assumed that by the end of the year, control children will look more like the low-SES children and Head Start children more like the middle-SES children. Thus, the difference in means for the two groups in the normed data is equivalent to an estimate of the Head Start effect. Standard deviations are similar across SES groups. An estimate of the common error variance for each measure is obtained by taking the mean of the standard deviations for low-SES and middle-SES children.

For each sample size per cell, Table 7-4 shows how power varies with differences in effect size (difference between low-SES and middle-SES children versus difference between low-SES and high-SES children) and different common error variances (mean of the variances for the three SES groups--low, middle, high--versus two-thirds of the mean of the variances for the three SES groups). Thus, Table 7-4 shows the effect on power of a larger treatment effect and the reduced within-stratum variation that could be obtained by stratifying the sample and by pretest and other background information.

As Table 7-3 and column 1 of Table 7-4 indicate, statistical power does not reach 0.50 for any of these measures until n = 25. If Head Start "jumps" the child two SES steps or if stratification reduces within-stratum variations (columns 2 and 3, Table 7-4), power is > 0.50 at n = 20. Power of 0.5 means that there is a 50 percent chance of rejecting the null hypothesis when the alternative is true.

Optimum Allocation. The previous section illustrates how to choose the minimum sample size of children per site for a designated precision. Tables 7-3 and 7-4 assume one dependent variable per calculation and an unstratified set of children. However, the evaluation involves:

- Multiple strata of children, probably with different mean differences and variances on the same dependent variable.
- Multiple dependent variables.
- Different attrition rates in different strata.
- Different costs per observation for different strata.
- Alternative budget levels.

- Alternative ceilings for manageable field operations.
- Different information priorities for different strata.
- Different optimal sample sizes for different sampling units (children and sites).

Under circumstances of variations between strata in the factors that affect the selection of sample sizes, *optimum allocation* is a way to achieve optimum precision for a given cost, or, conversely, lower cost for a given precision. Since numerous variables are relevant to determining the optimal distribution of sampling units among strata, it is almost necessary, and certainly preferable, to solve the problem with a formal allocation model. Conlisk and Watts (1969) developed the allocation model for determining cost-effective sample sizes in income maintenance experiments. Morris (1973) extended the model to treat selection from a finite population (e.g., Head Start-eligible children) with an infinite number of differences.

These sophisticated techniques work well with quite accurate information about the relevant variables. Unfortunately, fairly precise information on these variables is not available for the Head Start case. The pilot test of the evaluation (see Chapter 9) should supply more accurate estimates of the necessary parameters.

It should be remembered that optimum allocation offers economic advantages and increased field manageability. A random design with a larger number of children per site and number of sites per stratum provides precise estimates for system and stratum effects as an optimal design with a smaller sample. However, there are budget constraints and upper limits on the number of children per site and number of sites that can be managed well in the Head Start case. Thus, the contractor might consider using the results of the pilot evaluation to determine an optimum allocation of children among sites and sites among strata. It is advised that this work be subcontracted, since the models are new and, to be used properly, require experience with them and a high degree of computing sophistication. If it is intended to use optimal allocation for the final selection of sample sizes, the subcontractor should work with the contractor prior to the pilot evaluation to assure the best possible estimates for the model.

Recommended Sample Sizes

In the absence of a more precise way to determine appropriate sample sizes for children per site and sites per stratum, we recommend sample sizes by assuming certain constraints. On the assumption that two Head Start classes per site are the maximum expected number of classes and that there are 15 children per class, we recommend sampling two classes of Head Start children and an equal number (approximately 30) of control children per site.

The decision about number of sites per stratum for the site-by-site analysis is heavily constrained by field manageability and budget considerations. There is an upper limit on the number of sites that can be managed well and financed properly. The management and financial upper bounds are not necessarily the same. Experience in past Head Start evaluations indicates that a maximum of about 100 sites can be managed well. Cost estimates supplied to the OCD indicate that it will cost about $5,000 to administer the pretest per site to 60 children and $5,000 to administer the posttest per site for 60 children in the Head Start year. These numbers are exclusive of contractor overhead costs, the subsample and special sample studies, and the collection of data on the independent variables. Thus, the cost per site for 60 children of pretests and posttests is $10,000; for 100 sites, $1 million. The financial upper limit for the evaluation will have to be determined by OCD. Now the problem is the allocation of (let us assume 100) sites to different strata. If we assume 10 strata, and divide sites equally among them, there are 10 sites per stratum. The sample size for a site-by-site analysis of stratum effects is very small. Its adequacy in part depends on the variation in effect among sites. If there is considerable variation, the estimate of stratum effects will be imprecise. There are several choices:

- o Accept imprecise estimates. This means that some information is obtained about the properties of the distribution of Head Start effects for each stratum. In Bayesian terms, the information gained would serve as a diffuse prior. This information would still be of benefit, particularly for strata for which we have very little information (Chicano, Puerto Rican, Native American).

- Do not estimate effects for all strata. This means eliminating certain groups from the evaluation, possibly evaluating them later in a focused study. This is a policy choice that OCD is much better able to make than Rand.
- Allocate sites unevenly to obtain more precise estimates of some strata at the expense of others. Two bases for making these choices are: (1) more precise estimates for strata for which we know the least; or (2) more precise estimates for strata that represent the largest proportions of children served by Head Start. Since we know least about groups that represent small proportions of the total number of children served by Head Start, these criteria have opposite implications for allocating sites to strata. If certain strata have high within-stratum variance relative to other strata, unequal allocation may also be used to obtain equally precise estimates of all strata.
- Combine strata--e.g., pool sites for the Chicano and Puerto Rican strata. This is not an appealing solution if the recommended stratification is successful in creating more homogeneity within strata. Combining strata in this way simply reintroduces the variation. If the recommended stratification is not successful in these terms, combining strata is sensible.

Since there are management and financial constraints and we do not know within-stratum variabilities and OCD information priorities, the best solution is an equal number of sites per stratum for a total of 100 sites. If the 10 strata are retained, this implies 10 sites per stratum.

Chapter 8

ISSUES OF STATISTICAL ANALYSIS

CONFIRMATORY VERSUS EXPLORATORY DATA ANALYSIS 319

HYPOTHESIS TESTING VERSUS CONFIDENCE INTERVALS 322

FORM OF THE HYPOTHESIS 323
  One-Tailed Versus Two-Tailed Test of H_1 323
  Form of H_1 324

LEVELS OF CONFIDENCE AND SIGNIFICANCE 328

INTERPRETATION OF THE RESULTS 329

MODELS FOR THE ANALYSIS OF RANDOM ASSIGNMENT AND VALUE-ADDED DESIGN 330
  Random Assignment Design 330
  Statistical Analysis Strategy for a Value-Added Design 336

AGGREGATION OF MEASURES 337

Chapter 8

ISSUES OF STATISTICAL ANALYSIS[1]

In a standard laboratory experiment the model for the experimental situation structures both data collection and analysis. The analysis is thus fairly straightforward. However, as indicated in the Introduction to Chapter 7, the design for this evaluation is *generalized* to handle a *set* of research questions, no one of which should necessarily be analyzed by the statistical model implied by the design. In addition, a laboratory experiment and the national evaluation of a social program are very different research situations. The constituents to which the analysis is accountable are singular in a laboratory study and multiple in an evaluation, the costs associated with errors are different and so on.

This chapter discusses issues of statistical analysis that emerge as a result of those differences. The point is not to specify particular statistical descriptors and tests. Certainly some are consistent (and some are not) with the sorts of statements and inferences the research questions require and with properties of the data. However, we are confident that if the evaluation occurs, the analysis group will be as familiar as we are with alternative procedures. They will be more familiar with the properties of the data and the statistical implications of alternative criteria for success of the program that emerge in the course of the evaluation. Here we address issues of analysis and reporting that are not standard in an ordinary experiment--that is, *how to think about* the analysis of evaluation data and its communication to the policy world. In some cases we feel strongly about a particular perspective. In others we simply suggest that the analysis group consider particular ideas in working out their analysis plan.

The chapter discusses the following issues:

---

[1] We would like to thank Pierce Barker, Stephen Carroll, David Kenny, William Rogers, and Ralph Strauch of The Rand Corporation for their helpful contributions to this chapter.

- Confirmatory versus exploratory data analysis.
- Hypothesis testing versus confidence intervals.
- Form of the hypothesis.
- Levels of significance and confidence.
- Interpretation of the results.
- Models for the analysis of random assignment and value-added designs.
- Aggregation of measures.

CONFIRMATORY VERSUS EXPLORATORY DATA ANALYSIS

John Tukey (1970, Vol. I) distinguishes confirmatory and exploratory data analysis: "Exploratory data analysis is detective in character. Confirmatory data analysis is judicial, or quasi-judicial in character."[1] As indicated in Chapter 1, a national evaluation of a social program is most appropriate in judicial, or decisionmaking, situations, specifically in the context of legislative or budgetary decisions about the program. It is not an optimal research situation for generating clues, primarily because the usually requisite sample size restricts the number of variables that can be observed for any one unit in the sample. In a national evaluation, variation is usually a problem to be eliminated, not a source of discovery. In other words, if the information sought about a program is primarily generative, research resources are better allocated to small, intensive studies, not to a national evaluation.

As also indicated, of the four questions asked of Head Start,[2] only the question about effects (questions 1 and 2) warrant a national evaluation. We do not know enough about questions 3 and 4 to treat them judicially, although evaluation data should shed light on them.[1] Thus, if a national evaluation of the Head Start program occurs, we assume that the objective is to collect data for legislative or budgetary decisions about the program. Questions 1 and 2 should therefore be subjected to a statistical analysis consistent with the judicial character of the evaluation--i.e., to confirmatory data analysis. Questions 3 and 4, which are not optimally investigated by a national evaluation, in general can only be subjected to exploratory data analysis.

---

[1] The concepts of confirmatory and exploratory data analysis correspond to the concepts of *a priori* and *a posteriori* data analysis in the experimental design literature.

[2] The four questions are: (1) What are the social competence effects of Head Start for members of the eligible population who receive the treatment, relative to members of that population who do not? (2) What are the social competence effects of Head Start for eligible children from different cultural groups who receive treatment, relative to eligible children from those same groups who do not? (3) What are the social competence effects of Head Start for eligible children within each

The confirmatory and exploratory concepts have implications both for the *analysis of data* and for the *confidence that can be placed in the results*. With regard to data analysis, the distinction between confirmatory and exploratory analysis has at least three implications. First, in confirmatory analysis statistical methods are used to test previously generated hypotheses, not to generate new ones. In other words, tests of significance and confidence intervals are important. Second, the data are examined in confirmatory analysis, not for the hypotheses they might imply, but for their consistency with extant hypotheses; comparisons should be planned, or *a priori*, not *a posteriori*. Third, in the exploratory stage we are concerned with which hypothesis *might* be true. There is no interest in generating false hypotheses, but since generated hypotheses have a tentative status, errors are less important. In the confirmatory stage errors are more important, particularly in the context of a national evaluation. It may be true, as Rozeboom (1971, p. 120) argues, that "the primary aim

---

cultural group who receive the treatment and who differ in entry characteristics, as indicated by pretests and other background characteristics? (4) Are there any indications that variations in treatment produce variations in social competence outcomes for children who receive the treatment?

[1] This statement is true if: (1) in the pilot evaluation the recommended process variables do not successfully differentiate programs; or (2) even if the process variables successfully differentiate programs, the cost of double-sampling to establish a sampling frame for programs of different types is considered too high. If we are able to differentiate programs and sample explicitly to evaluate the effects of program variations, it becomes appropriate to treat question 4 judicially.

of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested." However, basic research is more of a "world enough and time" situation. That is not the case for a national evaluation, which is undertaken *in order* to "precipitate decisions." Since statistical decisions have direct implications for policy decisions, errors are important. In other words, in confirmatory analysis, particularly confirmatory analysis for social policy, being wrong matters. This means greater concern with the probabilities of Type I and Type II errors and with levels of confidence.

With regard to the confidence one can place in the results of confirmatory versus exploratory analysis, obviously statistically and substantively incorrect decisions can be made in either type of analysis. However, *a posteriori* analyses are more subject to both kinds of errors than *a priori* analyses. Exploratory analyses tend to be multiple, nonorthogonal, and conducted for *observed*, rather than predicted, large differences. These three reasons increase the probabilities of Type I errors for the set of analyses. There are statistical solutions to this problem that set the error rate for conceptual units larger than the individual comparison (e.g., for the experiment). However, these same solutions are less apt to detect real differences.

There is a tendency, in even the most rigorous of scientists, to forget the differential trustworthiness of *a priori* versus *a posteriori* results. Confidence tends to be more related to the consistency or inconsistency of the research results with the consumer's expectations or desires than to the *a priori* or *a posteriori* status of the results. In reporting the results, therefore, the analysts must specify how much trust the consumer of the research has a right to place in any given result.

The remainder of the chapter addresses considerations of confirmatory analysis, issues that primarily affect questions 1 and 2, since it is only for these two questions that an *a priori* analysis plan is indicated. With regard to the exploratory analysis of questions 3 and 4, our major concern is that it be systematic. We strongly advise that Tukey's volumes on *Exploratory Data Analysis* (1970) be used to structure the exploration. These volumes detail organized and creative ways to search. An example of the use of some of these techniques appears in Bunker et al. (1969).

HYPOTHESIS TESTING VERSUS CONFIDENCE INTERVALS

Both significance tests and confidence intervals involve the same assumptions. Confidence intervals provide more information than significance tests. They test any null hypothesis specified in the hypothesis-testing procedure and all possible null hypotheses. If the value specified by the null hypothesis occurs within the interval, the null hypothesis cannot be rejected. Confidence intervals also provide information about the magnitude of differences between two populations and about the error variation associated with an estimated difference. Significance tests tell us the probability that a true difference exists between two populations.

Confidence intervals alone are often used in reporting the results of policy research. As indicated above they convey more information to the statistician than a significance test. In one respect they convey more information to the policymaker. The appropriateness of a confidence interval for policy reports resides primarily in the information it conveys about the range of the magnitude of difference. It there is a difference between two groups, it is difficult to know from a P value what that difference "means." In significance testing the null hypothesis may be rejectable at $P = 0.10$. If 0.10 is greater than the Type I error specified for the study, the decision is not to reject the null hypothesis. However, we do not know what difference *might* exist between the two populations. Thus, even though we are unable to reject the null hypothesis on the basis of this sample, we cannot tell from a P value whether or not the difference that might exist on the basis of repeated samples is worth knowing. Similarly, if there is a large difference between, for example, the mean heights of the samples from two populations, this difference is expressed as a small P value. It is more meaningful to know that for a confidence limit at the 0.99 level the difference is, for example, 2" ± 1.57", or 3.57" to 0.43".

In one respect, hypothesis tests are more isomorphic with policy concerns than confidence intervals. Rozeboom (1971, p. 126) states conditions under which hypothesis testing is useful:

> While a confidence-interval analysis treats all the alternative hypotheses with glacial impartiality, it nonetheless frequently occurs that our interest is focused on a certain selection from the set of possibilities. In such cases, the statistical analysis should report, when computable, the precise p value of the experimental outcome.

A national evaluation is a confirmatory or decisionmaking exercise, not an exploratory exercise. Thus, if a national evaluation is conducted, it can be assumed that there are a limited set of decisions on which it is expected to impinge. Policymakers are interested in the implications of the data for this limited number of decisions, not for an infinity of decisions. In this situation it is preferable to test those hypotheses about the program that connect directly with the decision alternatives available to policymakers.

In sum, we argue that analyses should be reported both in terms of confidence intervals *and* as tests of hypotheses.

FORM OF THE HYPOTHESIS

The hypothesis-testing procedure works with two hypotheses: the hypothesis that is tested and the alternative hypothesis. The hypothesis that is tested is usually called the "null" because the form is frequently a statement of "nothing," or no difference. This is not a necessary form, and consequently we prefer to refer to the "null" and "alternative" hypotheses as H_1 and H_2. There are two questions involved in the form of the hypothesis that is tested--i.e., H_1. One is the question of whether we want H_1 and H_2 to exhaust all logical possible outcomes of the experiment (i.e., to involve both tails of the distribution). The second is the form we want H_1 to take. These questions are discussed in turn below.

One-Tailed Versus Two-Tailed Test of H_1

The logical, possible outcomes of the experiment are: (1) $HS > \overline{HS}$;

2) $HS = \overline{HS}$; and (3) $HS < \overline{HS}$. In some research situations, if there is a difference between two treatment levels, we expect that it will be in one, rather than the other, direction. In other cases we may have no basis for predicting the direction of difference but are only *interested* in one direction. In either of these two cases it is preferable to exclude one tail of the distribution from the test of H_1, since a one-tailed test is more powerful than a two-tailed test.

In the Head Start case the usual results for the variables measured have been either $HS = \overline{HS}$ or $HS > \overline{HS}$.[1] However, this evaluation proposes outcomes not previously measured. For these we are unable to predict the direction of difference, should one occur. In addition, we *are* interested in both directions: $HS > \overline{HS}$ has very different implications for social policy than $HS < \overline{HS}$.

In sum, the form of H_1 and H_2 should exhaust all logical, possible outcomes of the experiment: $HS = \overline{HS}$; $HS > \overline{HS}$; and $HS < \overline{HS}$. In other words, they should cover all possibilties in the parameter space for the distribution of one or multiple variables.[2]

Form of H_1

Although the traditional two-tailed form of H_1 and H_2 is $HS = \overline{HS}$ versus $HS \neq \overline{HS}$, H_1 and H_2 can take any of the following two-tailed forms:

1. H_1: $HS - \overline{HS} = k$, where k is a specific number, including 0;[3]
   H_2: $HS - \overline{HS} \neq k$;

---

[1] HS = treatment condition (Head Start); $\overline{HS}$ = control condition (no Head Start).

[2] The same argument for the form of H_1 and H_2 applies to the use of confidence intervals. We recommend a two-sided confidence limit for comparisons of means or proportions.

[3] The definition of k depends on what is defined as an effect for the evaluation. It can be a statistically significant difference, in which case $k = 0$. It can be a number identified as a "policy-significant" difference--e.g., k = the difference between middle-class and lower-class children for a normed measure; k = a proportion of children who reach a particular threshold. Obviously, implications of the same data for different definitions of k can be included in the analysis.

Final definition(s) of effect should be selected after the evaluation and before the analysis. Scientifically, it is preferable to state the definition of effect before the data collection. However, the major

2. H_1: $HS - \overline{HS} \geq k$
   H_2: $HS - \overline{HS} < k$;
3. H_1: $HS - \overline{HS} \leq k$
   H_2: $HS - \overline{HS} > k$;
4. H_1: $HS - \overline{HS} > k$
   H_2: $HS - \overline{HS} \leq k$;
5. H_1: $HS - \overline{HS} < k$
   H_2: $HS - \overline{HS} \geq k$.

Although we recommend a form, our major concern in this section is to detail the reasoning that led us to the particular recommendations. At the time of the data analysis, the policy environment may have changed considerably. In that case a different form of H_1 and H_2 may be indicated. Considerations for that choice are specified here.

Neyman (1942) suggests a criterion for selecting the form of H_1: Equate with the statistical hypothesis (i.e., H_1) that empirical (i.e., substantive) hypothesis for which the error of erroneous rejection is more serious than the error of erroneous acceptance.[1] In this case the

---

concern is with definitions that are meaningful. Policy-significant definitions are useful only if they reflect the policy environment current during the analysis phase. We also expect candidate definitions to surface during the conduct of the evaluation from Head Start parents and personnel. Both of these considerations militate against final decisions on definitions of effect before the conduct of the evaluation.

[1] The question about accepting the null hypothesis places us in the middle of a philosophical debate. Fisher (1949) argues that the "null hypothesis is never proved or established, but is possibly disapproved, in the course of experimentation." As other statisticians have pointed out, this is not a very helpful statement, since one does not expect to prove any hypothesis by the methods of probabilistic inference. The possible parameters for the distribution of a random variable are conceptually represented in a parameter space. If we conceive of this space as divided into two subsets, H_1 specifies that the parameter occurs in subset$_1$; H_2, in subset$_2$. Failure to reject H_1 is the same as deciding that the parameter occurs in subset$_1$. The problem with talking about accepting H_1, of course, is that the hypothesis is stated as either a point or interval estimate. The point by definition is bounded--e.g., 0 or 4 or 12.5. The interval can be bounded at least at one end--e.g., ≥ 0 or 4 or 12.5. However, any subset of the parameter space includes not only points specified by the hypothesis, but also points consistent with the hypothesis under assumptions about sampling error.

more important error is under direct control of the investigator. If we use this criterion we have to determine: (1) the set of possible outcomes of the experiment for each research question, (2) the policy decisions at issue for each research question, (3) the correspondence of alternative decisions to alternative outcomes of the experiment, (4) the correspondence of alternative outcomes to alternative forms of H_1, and (5) the risk of error associated with each alternative decision and consequently with its corresponding form of H_1.

Let us take research question 1: What are the social competence effects of Head Start for members of the eligible populations who receive the treatment relative to members of that population who do not? As indicated above, there are three basic outcomes for this question for any one indicator or set of indicators of social competence: (1) $HS > \overline{HS}$; (2) $HS = \overline{HS}$; and (3) $HS < \overline{HS}$. For constituents with federal budgetary control over the Head Start program, there are three basic actions, or decisions, they can take with regard to the Head Start program: (1) increase the budget for the program, (2) maintain the budget,[1] and (3) cut the budget. In the current economic environment, option 1--increase the budget--is improbable, regardless of the demonstrated value of the program. Thus, the two decisions more probably at issue are maintain the budget and cut the budget.

The correspondence between alternative probable policy decisions about the program and outcomes of the experiment would seem to be: maintain the budget $\equiv HS > \overline{HS}$ or $HS = \overline{HS}$; and cut the budget $\equiv HS < \overline{HS}$. While the budget would improbably be cut under a $HS > \overline{HS}$ outcome, the correspondence between budget maintenance and $HS = \overline{HS}$ is more questionable. The argument, plausible in this policy environment but not necessarily in a changed environment, is that there is a strong political constituency for the Head Start program. Given this constituency, we feel that it would take more than a "no difference" finding in an impact evaluation to cause a budget cut in the Congress, OMB, or other Executive offices with budgeting power over the Head Start program.

---

[1] "Maintenance" can include budgetary increases to offset cost measures as the result of inflation.

This correspondence between probable policy actions and outcomes of the experiment implies two alternative forms of H_1: (1) $HS - \overline{HS} \geq 0$; or (2) $HS - \overline{HS} < 0$. If we return to Neyman's criterion for selecting the form of H_1, the final question is: For which of these two statements is the error of erroneous rejection more serious than the error of erroneous acceptance? The decision is clearly a value, or normative, decision, informed to some extent by empirical knowledge. We take a position here but recognize that other constituents to the evaluation can have different positions. We argue that $HS - \overline{HS} \geq 0$ should be chosen as the form of H_1, in the belief that erroneous rejection of this statement is more damaging than its erroneous acceptance. The reasoning is as follows. Erroneous rejection of $HS - \overline{HS} < 0$ is more serious than erroneous rejection of $HS - \overline{HS} \geq 0$ if we have reason to suspect that HS is actually harming children. To our knowledge there is no reason to expect this Head Start effect. It *is* possible that HS is not accomplishing as much for eligible children as $\overline{HS}$ alternatives. To fail to discover this costs us the opportunity to allocate funds optimally among preschool alternatives for poor children. This is a cost. However, the *opportunity to discover that $HS - \overline{HS} < 0$ remains unaffected*. In other words, there are other chances to discover the error.

Erroneous rejection of $HS - \overline{HS} \geq 0$ carries the probable consequences of cut-back or cancellation of the program. There are three costs associated with this error, one of them contingent on what is done with the released funds. First, *if* $HS > \overline{HS}$, there are opportunity costs for eligible children even if released funds are placed into $\overline{HS}$ preschool alternatives. Second, even if HS is accomplishing no more than $\overline{HS}$ alternatives, there are opportunity costs if released funds are not invested in other programs for poor children. Third, if erroneous rejection of $HS - \overline{HS} \geq 0$ results in cancellation of the program, the opportunity to discover that $HS - \overline{HS} \geq 0$ is lost. In other words, there will be no chance to discover the error.

In sum, readers may disagree with our reading of probable policy decisions and with the costs we assign to alternative errors. However, the reasoning here represents a mechanism by which to make a statistical choice that is isomorphic with the policy decision objectives of a national evaluation.

LEVELS OF CONFIDENCE AND SIGNIFICANCE

Levels of confidence and significance are under the direct control of the analyst. Although these levels both involve the concept of error, the interpretation of error differs for confidence interval and significance test procedures. A 99 percent confidence level and a significance level of 0.01 both designate a 0.01 probability of error. However, for confidence intervals error is the expected proportion of intervals that do not contain the true population parameter--in this case, 1 percent of the intervals. For tests of hypotheses error is the proportion of hypotheses erroneously rejected. The point of this section is to discuss criteria for selecting levels of error for the Head Start evaluation.

We expect confidence intervals to be more relevant to analysts than to decisionmakers. For analysts the costs of error tend to be the costs of lost information. The question is then: How does information vary as a function of variation in expected error? A lower confidence level narrows the limits placed on a sample estimate. Thus, it reduces the alternative values the true population difference is expected to assume. It also is less apt to include the true population difference. If "information" is defined as reduced uncertainty, a lower confidence level increases information on one dimension and decreases it on the other. The costs and benefits of a high confidence level are the reverse. Our preference is to select high confidence levels because they define the "maximum region" in which we can expect the true population difference to occur.

The selection of significance levels is more serious because we expect tests of hypotheses to link more directly to policy decisions than confidence intervals. Since there is an inverse relationship between Type I and Type II errors, selecting high protection against a Type I error reduces protection against a Type II error. If the form of H_1 is chosen on the basis of Neyman's criterion, there is some social protection against the costs of a Type II error. Thus, the tradeoff between the two types of errors should be in favor of minimizing the probability of a Type I error.

The rational basis for choosing an α level is the loss function associated with the two types of errors. Neyman used the *concept* of a loss function as the basis for choosing the form of H_1--and that basis has implications for choice of an α level. However, using the concept of loss function for thinking about Type I and Type II errors is different from calculating such a function. To calculate optimal values of α, it would be necessary to identify and quantify the costs of Type I and Type II errors. In social policy it is exceedingly difficult to imagine all possible consequences of making each type of error. Even for those consequences that can be envisioned, it is difficult to assign credible numerical values for their costs.

In the absence of being able to calculate a loss function for Type I and Type II errors, there are at least three different bases for selecting an α level. First, we can take the conventional definition of a conservative α level--e.g., 0.05 or 0.01. The advantage of this procedure is that it is conventional; the disadvantage, that it is arbitrary. Second, we can solve for the P value for each hypothesis--i.e., the value of α at which the decision regarding H_1 is on the borderline between acceptance and rejection. This procedure allows the readers to apply their own standards of tolerable error. The disadvantage is that no single point is designated as the rejection or not-rejection point for H_1. Third, we can select conservative and less conservative P values--e.g., 0.01 and 0.10--as the rejection and acceptance points for H_1. P values between these two points represent the range for consumer judgment. For example, if the P value for a particular test = 0.06, the consumers of the results can decide whether they are willing to reject H_1 at this error level. This procedure combines the advantages of the first two procedures.

INTERPRETATION OF THE RESULTS

It is important to remember what different outcomes of a HS *versus* $\overline{\text{HS}}$ experiment can mean. In the case of HS > $\overline{\text{HS}}$, we can conclude that HS is doing better than its competitors together, but not necessarily better than any one competitor. During the Head Start year, control children are in a variety of home care, informal day care, and formal day care situations. Unless we can demonstrate that these different situations have the same effects on the child, we cannot conclude that HS is better than *each* of these control treatments.

In the case of HS = $\overline{HS}$, we cannot conclude that HS has no effect. We can only conclude that HS has the *same* effect as the combined set of control treatments. Many control children could be enrolled in excellent day care programs in their communities. In this case, a finding of no difference can simply mean that both the Head Start and day care program are helping children.

In the case of HS < $\overline{HS}$, we cannot conclude that HS is "harming" children. We can only conclude that HS is having less effect on child outcomes than the combined set of control treatments.

MODELS FOR THE ANALYSIS OF RANDOM ASSIGNMENT AND VALUE-ADDED DESIGNS

The random assignment design and value-added design are consistent with different analysis models. These issues are discussed for each design in turn. The discussions are limited to research questions 1 and 2 since, as indicated, an *a priori* analysis plan is primarily relevant to these two questions.

Random Assignment Design

The random assignment design for data collection was presented in Table 7-1. That design is consistent with several different analysis models, depending on the amount of information we want to incorporate in any single test. That decision depends on substantive and statistical considerations.

As indicated, the design for the evaluation is a generalized design for collecting data required by the total set of research questions. The statistical model behind the design is the general linear model, specifically an analysis of variance (ANOVA) model. In ANOVA terms, the design represents a randomized block partial hierarchical experiment, with sites nested in strata. However, even research questions that require all observations within the design--e.g., question 1--do not have to be analyzed according to the randomized block partial hierarchical model. For example, question 1 requires an estimate of the overall effect of Head Start. Thus, it does not *require* an analysis model that distinguishes either sites or strata. A randomized block partial hierarchical design, randomized block design, or completely randomized designs are all consistent with the substantive requirements of question 1. Question 2 requires an analysis of observations within each stratum separately but, again, does not require an analytic model that distinguishes sites within each stratum. Randomized block or completely randomized models, but not a randomized block partial hierarchical model, are consistent with the analysis requirements of question 2. Thus, there is a choice of ANOVA models for questions 1 and 2. There is also a choice between a cross-site and within-site unit of analysis. In the within-site case, cross-site estimates are achieved by aggregating within-site results by other statistical procedures. The decision about unit of analysis effects the decision about ANOVA models. Table 8-1 presents the ANOVA analysis alternatives for questions 1 and 2.

For evaluating questions 1 and 2, we recommend estimating effects for individual sites and aggregating site effects. If a variable--e.g., age--is found in the pilot test to vary sufficiently within-site to warrant its use as a blocking variable, we recommend a randomized block design for analyzing within-site observations.[1] If no such variable emerges, we are dealing with a completely randomized design for two independent samples. In this case a test for the comparison of means of two samples, not the F test, is appropriate. Procedures for aggregated site results across strata or within a stratum are discussed below.

There are substantive and statistical reasons for this recommendation. The strategy implies treating each site as a separate experiment

---

[1] If an "effect" is defined as difference in proportions of treatment and control children who reach a particular threshold, the measurement properties of the data shift from interval to attribute. In this case the ANOVA model is inappropriate. A test of a difference in proportions for one-way or two-way classifications (comparable to the completely randomized design or randomized block designs) is appropriate.

Table 8-1

ANALYSIS OPTIONS FOR QUESTIONS 1 AND 2

Type of Experimental Design	Mathematical Model	Observations Used in Analysis			
		Question 1		Question 2	
		All Observations	Observations Within Site	All Observations Within Stratum	Observations Within Site
Randomized block partial hierarchical design	$X_{ijkn} = \mu + \alpha_i + B_j$ $+ \gamma k(j) + \alpha B_{ij}$ $+ \alpha\gamma ik(j)$ $+ \varepsilon_{n(ijk)}$ (mixed model)	Design uses all data and estimates these effects: α_i = treatment effect B = stratum effect γ = site effect αB = interaction effect of treatment and stratum $\alpha\gamma$ = interaction effect of treatment and site	Not relevant--site and stratum are not treated as variables in the design.	Not relevant--stratum is not a variable for this question.	Not relevant--stratum is not a variable for this question.
Randomized block design	$X_{ikn} = \mu + \alpha_i + B_k$ $+ \alpha B_{ik}$ $+ \varepsilon_{ikn}$ (mixed model)	Design uses all data and estimates these effects: α = treatment effect B = site effect *or* stratum effect αB = treatment/site *or* treatment/stratum interaction If B = site effect, stratum is not treated as a variable in the design. If B = stratum effect, site is ignored as a variable, i.e., analysis is conducted on observations aggregated across sites.	Not relevant unless a within site blocking variable is introduced, e.g., age. In this case design estimates for a site: α = treatment effect B = age effect αB = treatment/age interaction effect Treatment effects/site are aggregated across sites and strata by other procedures.	Design estimates for stratum$_j$: α = treatment effect B = site effect for all sites in stratum$_j$ αB = treatment and site interaction effects	Not relevant unless a within-site blocking variable is introduced, e.g., age. In this case design estimates for a site: α = treatment effect B = age effect αB = treatment/age interaction effect Treatment effects/site are aggregated across sites within stratum by other procedures.
Completely randomized design	$X_{in} = \mu + \alpha_i$ $+ \varepsilon_{in}$ (mixed model)	Design uses all data and estimates this effect: α = treatment effect Analysis is conducted on observations aggregated across sites and strata.	Design estimates: α = treatment effect for a site. Treatment effects/site are aggregated across sites and strata by other procedures.	Design estimates: α = treatment effect for a stratum. Analysis is conducted on observations aggregated across sites within a stratum.	Design estimates: α = treatment effect for a site. Treatment effects/site are aggregated across sites within stratum$_j$ by other procedures.

and the sample of sites as a sample of experiments. Analytic procedures that *simultaneously* use all observations relevant to the evaluation of a research question are more efficient. The problem with these procedures lies in the *meaningful aggregation of observations*. Although we discuss this in ANOVA terms, it is a problem regardless of specific technique.

Questions 1 and 2 require estimates of treatment effect. Obviously, it is important that these estimates have meaning within the context of the program. *A single analysis based on all observations relevant to either question 1 or question 2 will only improbably yield a meaningful estimate of effect.* Previous Head Start analyses indicate substantial intersite variation in effects--children, communities, treatments, or all three vary. Although this assumption should be checked in the pilot data, we expect to find the same intersite variation in this evaluation. In ANOVA terms intersite variation emerges as site/treatment or stratum/treatment interactions. In other words, the treatment has different effects for different children--or, since the variations in treatment almost certainly make a difference in outcomes for children, different treatments are having different effects on the same children. A significant interaction term nearly always requires a qualified interpretation of the main effect of treatment and can render it uninterpretable. We cannot say that Head Start has an effect of x amount on children because in fact it has different effects for different children. Sources of the interaction can be determined. If there is *patterned* interaction, it may be possible to talk about a main effect in a qualified way. For example, we might be able to state that there is x effect of randomly selected central city Head Start centers on Chicano children. However, we cannot count on being able to interpret the interaction term, particularly given the inability to stratify treatments.

The model for completely randomized design does not include an interaction term since neither stratum nor site is a variable in the design. Although this design does not involve the interaction problem statistically, it only sidesteps substantively. It also creates a statistical problem in detecting treatment effects: The variance contributed by stratum and site dimensions are pooled with the error term.

If intersite variation is large, the size of the error term is apt to obscure genuine effects.

A solution to the problem of a meaningful estimate of treatment effects is to estimate effect for each site separately. We argue that the results for an individual site are interpretable. *They represent the effects of the program as it is modified by community characteristics and community-specific needs of the children.* We also argue that an aggregation of site-specific results is interpretable as an estimate of overall effects of the Head Start program. In evaluation design there is a tendency to think about local manifestations of a national social program as replications of an experiment. Intersite variation in the program is then treated as unhappy deviation from a replication model. We argue that the replication model is not appropriate. In major social programs--Head Start, Title I, the Manpower Program--local options are *mandated* as part of the program. Variation in implementation then becomes *part* of the experiment. At the same time all instances of the program are accountable to certain across-site objectives. *The appropriate estimate of national program effect then becomes an aggregation of the effects of the program for measures of those objectives, as they are realized at the local level.*

Estimating program effects by aggregating individual site effects has two major statistical consequences. A within-site comparison between treatment and control groups has many fewer degrees of freedom than a comparison based on cross-site aggregations. This reduces the power of the test. However, as noted in Chapter 7, statistical power is a function not only of sample size but also of within-group variation. A within-site comparison holds treatment and community constant.[1] We expect that removal of these sources of variation in cross-site comparisons will gain more power than the smaller sample size will lose.

---

[1] Since the Head Start experience actually occurs in a specific classroom in a specific center, the classroom *versus* control would be the most precise comparison. However, as indicated, the sample size problem forces us to consider two classrooms. Although we can expect variation between classes, we can expect the treatment to be much more similar between two classes within the same center than between two centers.

In order to increase the precision of the within-site estimate, we also strongly recommend using pretest scores and other child and family background variables to remove sources of variation among children within a site before comparing T and -T scores. This suggestion implies an analysis of covariance technique.[1]

Estimates of effect for questions 1 and 2 by aggregating site-specific results require meaningful aggregation procedures. This is a point in the analysis when it is easy to select a statistically appropriate and substantively empty summing measure. Three suggestions for summing are as follows.

Site effects can be displayed in a frequency distribution, effects at each site being standardized by dividing the T/-T difference for each site by the pooled within-group standard deviation for that site. Properties of the distribution can then be assessed. Although it is possible to test for a statistically significant difference between the mean for the distribution and zero, other properties of the distribution may be more meaningful, e.g., skewness properties.

Another way to aggregate site effects is to define sets of increasingly stringent definitions of success, displaying the fraction of successful sites as a function of increasing stringency. The concept of increase can be interpreted in a variety of ways--e.g., an increasing number of measures on which sites can be successful or an increasing amount of improvement in an ability.

A third way to aggregate site effects allows us to assess the overall significance of the set of the differences between T and -T groups at each site. If we assume that statistically the sites are analogous to independent replications of an experiment, the statistical significance for the total set of site effects can be obtained. One technique is to convert the *P* levels for each site to a standard normal deviate, sum the deviates for the set of sites, take the square root of the sum and use the standard normal value to yield an overall test of significance.

---

[1] The recent statistical debate about the appropriateness of covariance techniques for adjusting differences between groups (e.g., Lord, 1960, 1967, 1969; Porter, 1967) refers to the use of ANCOVA for the correction of *systematic*, rather than *chance*, variation between groups. Nevertheless, analysis of covariance makes stringent assumptions about the data. It is important to check whether the data violate the assumptions, particularly the assumption that the regression of the dependent variable on the covariate is the same for all treatment populations.

Statistical Analysis Strategy for a Value-Added Design

Although there are aggregation alternatives for a random assignment design, the structure of the value-added design allows only the choice of a stratum or system-based growth curve. In a value-added design, comparison occurs between a treatment group and a statistically created control group. In order to obtain sufficient cases to create the growth curve, at least the total sample for a stratum has to be used. Thus, a within-site comparison is impossible: controlled for age, the scores of treatment children are compared with a mean value for a cross-site group. As indicated earlier, this strategy probably yields a less precise estimate of effect than a within-site comparison.[1]

We recommend calculating separate growth curves for each stratum. Since we would expect greater variation in scores at each age across strata than within each stratum, estimates of stratum effects by means of a stratum-based growth curve should be more precise. This recommendation assumes that each stratum provides sufficient data points for each relevant age to create a stable curve.

To estimate system effects (question 1), the results of the ten strata can be aggregated in one of the three ways described for a random assignment design.

The statistical strategy for conducting a value-added analysis is discussed in Smith (1973), Weisberg (1973), and Bryk and Weisberg (1974). The reader is referred to those papers for detailed descriptions of the techniques and for discussion of problems of estimating the growth curve itself, and using the curve to estimate the effects for treatment children.

---

[1] This evaluation provides an excellent opportunity to estimate the difference in precision of the two designs. For all random assignment sites, analyses both for a random assignment design and for a value-added design should be conducted and their differences evaluated.

AGGREGATION OF MEASURES

Decisions about aggregating measures are substantive. Measures should be aggregated only if together they provide a *meaningful* summary property of children. For example, it may make sense to aggregate all the CIRCUS language measures to assess the child's overall language competence. We have recommended that the contractor elicit from Head Start parents and personnel their priorities among the measured dimensions of social competence. If there is consensus about priorities within a constituency, these priorities might be treated as joint indicators of social competence.

Measures can be aggregated by constructing an index or a vector of variables. An index has fewer degrees of freedom than a vector approach. The estimate of the relationship between treatment and control groups for an index is a univariate statistical problem if the set of measures is aggregated into a single index; for a vector of variables, a multivariate problem. In the vector case, multivariate analysis of variance (MANOVA) is a technique for estimating the relationship between groups on a set of measures. If there are only two groups, as in the Head Start case (control and treatment), Hotelling's multivariate T^2 test is appropriate. Both an index and multivariate technique assess the joint contribution of the set of measures to the "closeness" or "distance" between groups. If this analysis yields a significant treatment effect, it means that the average Head Start child is located in a significantly different region of the outcome space than the non-Head Start child.[1]

In either aggregation strategy there is usually a problem in estimating the individual contribution of measures to the measurement model for the index or vector. It is important to understand the individual contributions of measures to that model in order to assess the substantive coherence of the set of measures. If components of an index or vector vary enormously in their *individual* contributions to that index or vector, it is difficult to interpret the set of components as a coherent measure of a multivariate concept such as "social competence" or "language competence." If measures within the set are independent of each other, the multivariate prediction function partitions the total variance explained by the set of variables into a set of independent components of variance, each due to one of the predictor elements. In other words, in the independent case it is fairly simple to estimate the individual contributions to the model for the index or vector. However, if measures within the set are correlated with each other--*and usually they are*--there is no entirely substantively satisfying solution to the problem of individual contribution. The objective in this situation is to construct an uncorrelated vector variable by an orthogonalizing transformation of the data. There are several statistical techniques for locating orthogonal dimensions--e.g., factor analysis, including the method of principal components; discriminate analysis, if groups are defined as the independent variable and the discriminate functions as the most predictable functions of the dependent variable vector. However, there are two major problems here, the first exacerbating the second. First, there are an infinite number of such transformations (i.e., mathematically, there is no unique solution). Second, we have to be able to interpret the orthogonal dimensions substantively in order to interpret the index or vector. The less we understand the theoretical relationship between variables in the measurement set, the more difficult it is to interpret orthogonal dimensions satisfactorily. The problem is less frequently in being able to make any interpretation than in making an inappropriate one. As Armstrong and Soelberg (1968) showed, expert judges are able to interpret even randomly generated factors, or dimensions. Thus, appropriate aggregation--whether of items for a single measure or of a set of measures--depends primarily on a substantive grasp of the variables involved and minimally on statistical manipulations.

---

[1] It is possible that there can be no significant difference between T and -T children for each measure individually, but there can be a significant difference between them on the collective measures. This sort of result is interpretable. If the measures are uncorrelated, the implication is that Head Start affects a child on any given variable only a small amount, but that the collection of small amounts places the Head Start child in a significantly different position in the outcome space than the control child. If the measures are somewhat correlated, one measure is picking up some, but not much, of the variance; together the measures are picking up a significant amount.

Chapter 9

TEST DEVELOPMENT AND PILOT TEST OF THE NATIONAL EVALUATION

INTRODUCTION 340

TEST DEVELOPMENT 341

COMPILATION OF THE STRATIFIED SAMPLING LIST 342

SELECTION OF THE SAMPLE OF SITES FOR THE PILOT EVALUATION 357
  Strata To Be Included 357
  Choice Between Probability and Purposive Sampling of Sites 358

PREPARATORY FIELD OPERATIONS 359
  Phase 1 of the Pilot Evaluation 359
  Preparation for the Full-Scale Evaluation 367
  Phase 2 of the Pilot Test of the Evaluation 374

Chapter 9

TEST DEVELOPMENT AND PILOT TEST OF THE NATIONAL EVALUATION

INTRODUCTION

This chapter provides a statement of tasks necessary for the national Head Start evaluation and of procedures for conducting a pilot test of the evaluation. It does not deal with the focused studies, which are different in intent and scale from the national evaluation and should be conducted independently of it.[1]

The full-scale evaluation covers two school years, 197X_1 - 197X_2 and 197X_2 - 197X_3. The activities preparatory for the full-scale evaluation also cover two years, 197X_0 - 197X_1 and 197X_1 - 197X_2. In other words, the preparatory period begins 12 months before the beginning of the full-scale evaluation and continues through the first year of that evaluation. Thus, if the contractor must enter local communities in July 197X_1, to solicit community support, arrange random assignment, etc., for the full-scale evaluation, the preparatory period should begin approximately in July 197X_0.

The main purpose of the pilot evaluation is to learn whether it is possible to protect the integrity of the full-scale evaluation, and if so, how. The criteria for integrity of the evaluation are data reliability and validity, ability to attribute differences between the treatment and control group to the treatment, and protection of the rights of local constituents, as specified in Chapter 1. *If the pilot evaluation indicates that these conditions cannot be met for the full-scale evaluation, the full-scale evaluation should not be conducted.*

This chapter does not present an exhaustive plan for the preparatory period. However, it does list major tasks that have to be accomplished during that time and indicates ways in which results of test development and the pilot test of the evaluation can be used to protect the integrity of the full-scale evaluation.

---

[1] Chapter 10 presents a detailed discussion of the focused studies.

The chapter is divided into four sections:

- Test development (adaptation, construction).
- Compilation of the stratified sampling list.
- Selection of the sample of sites for the pilot evaluation.
- Field operations from the winter of $197X_0$ - $197X_1$ to the spring of $197X_2$.

TEST DEVELOPMENT

As indicated in Chapter 2, Tables 2-1 and 2-2, several recommended tests need no development. Of the measures that have not been fully developed, the majority are scheduled for administration in the *post-Head Start year*. However, since some measures that need development work are scheduled for the pilot test of the evaluation in the spring of $197X_1$, we recommend that the contractor begin test development work immediately after the contract becomes effective. The development required ranges from a check of the clarity of instructions for the Head Start age group or for children from different ethnic groups to complete development of a new measurement idea.

We strongly recommend that the contractor subcontract development of each of the measures in Table 9-1 to groups knowledgeable about:

- The substantive area involved in the measure.
- The cultural context of responses of children, parents, and teachers from different SES and ethnic groups.
- The psychometric properties of tests.
- Reliability and validity problems specific to the type of instrument--e.g., archival retrieval, scale construction, observational schemes.
- Test administration considerations introduced by local testers.
- Management problems introduced by multiple site administration.

We also recommend that the contractor require the test developer to pilot the test with small samples of children for whom the reliability and validity of the test instructions and items might differ, to supply the contractor with the reliabilities of the test for each subgroup, and to discuss the validity of test items. The test developer should consider three types of validity: face validity, predictive validity (correlations between the scores on the candidate test and scores on tests with which we would expect a theoretical relationship), and scope condition validity[1] (specification of the conditions under which a test of an outcome is expected to be valid and demonstration that the particular test falls within those conditions).

If the reliabilities are low or validities questionable in general or for specific groups, *the test should not be included in the pilot test of the evaluation.* If the test is not administered in the pilot test of the evaluation, it should not be included in the full-scale evaluation unless other documented experience has proved it acceptable.

If an effort to develop a test fails or cannot be completed in time, there is a tendency to "redeem" the situation by including the test in the evaluation anyway, on the rationale of "trying it out." *A test should be included in the pilot evaluation and later in the full-scale evaluation if and only if there is documentary evidence that it will yield trustworthy data.*

For readers interested in more detail about which tests need to be developed and the type of development required, Table 9-1 provides a list of development requirements and a schedule of test administration in the pilot test and in the full-scale evaluation.

COMPILATION OF THE STRATIFIED SAMPLING LIST

Before a sample of centers can be selected, either for the pilot or for the full-scale evaluation, a list of centers has to be compiled, stratified by the ethnic, regional, and metropolitan or nonmetropolitan distinctions recommended in Chapter 7. For both random assignment and value-added designs the contractor needs the following information for

---

[1] For example, it might be argued that a valid test of object recognition requires that objects on the test be common ones in the child's environment. A content analysis of photographs of that environment might be used to demonstrate the frequency of objects in the environment and their consequent appropriateness for inclusion on the test.

Table 9-1

TEST DEVELOPMENT REQUIREMENTS AND SCHEDULE

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
I. Health/ nutrition	Treadmill or step test	RS	Adapt adult test for preschool children	End HS year, 197X_1	End HS year, 197X_2 (pretest in fall 197X_2 is optional)
II. Perceptual/ motor/ cognitive/ language	CIRCUS 2, 3, 4, 5, 13, and 16	W	Examine for appropriateness to Black, Spanish-speaking, and Native American children	End HS year, 197X_1	Beginning HS year, 197X_1
	CIRCUS 1, 9, and 10	SpS: English-speaking children	Examine for appropriateness to Black and Native American children	End HS year, 197X_1	Beginning HS year, 197X_1
	CIRCUS 1a	SpS: all Spanish-speaking children	Construct a Spanish equivalent of CIRCUS What Words Mean	End HS year, 197X_1	Beginning HS year, 197X_1

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
II. Perceptual/ motor/ cognitive/ language (cont'd)	CIRCUS 1b	SpS: all Spanish-speaking children	Construct a test of English vocabulary expected of children in English-speaking classrooms	End HS year, $197X_1$	Beginning HS year, $197X_1$
	CIRCUS 9a	SpS: all-Spanish-speaking children	Construct a Spanish equivalent of CIRCUS Listen to the Story	End HS year, $197X_1$	Beginning HS year, $197X_1$
	CIRCUS 10a	SpS: all Spanish-speaking children	Construct a Spanish equivalent of CIRCUS Say and Tell	End HS year, $197X_1$	Beginning HS year, $197X_1$
III. Social and personal development	Sociometric task: peer nominations; picture naming	W, (R_{rs})	Practically no work needs to be done	Winter $197X_1$-$197X_2$	Fall $197X_2$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Structured observation during free play (Ogilvie and Shapiro)	W, (R_{rs})	Some work--Step 1: Decisions must be made about exactly how to sequence observations given varying numbers of subjects per class and per school; Step 2: Good self-coding score sheets must be devised	Winter $197X_1$-$197X_2$	Fall $197X_2$
	Automatic time-sampling with camera	RS	The whole technique must be developed	Winter $197X_1$-$197X_2$	Fall $197X_2$
	Kelly role construct reportory test	W, (P_{rs})	Work aims at revising instructions and answer forms; namely, Kelly test (role constructs) needs instructions adapted for student roles in relation to parents and teachers, and an exact response format chosen from the existing alternatives	Winter $197X_1$-$197X_2$	Fall $197X_2$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Behavior rating	W	Content-wording and response scaling. Wording and scale intervals must be decided for: (1) summary estimates from teachers and parents and (2) early adjustment scales	Winter $197X_1$-$197X_2$	Fall $197X_2$
	Large item pool for CBI (e.g., the California Child Q Set)	RS	Little work needs to be done	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Semantic differential rating of picture stimuli	RS	Some work needs to be done	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Structured observation during an informal indoor task (Ogilvie and Shapiro, U.S. Commission on Civil Rights; Grotberg appendix)	W, (R_{rs})	Some work. Target behaviors chosen from Ogilvie and Shapiro for child-child and child-teacher interaction and from Bronson for executive skills. Executive skill categories need most work. Decisions must be made about exactly how to sequence observations given varying numbers of subjects per class and per school. Good self-coding score sheets must be devised	Winter $197X_1$ $197X_2$	Fall $197X_2$
	Kelly role construct repertory test	W, (P_{rs})	Work aims at revising instructions and answer forms. Kelly test (role constructs) needs instructions adapted for student roles in relation to parents and teachers, and an exact response format chosen from existing alternatives	Winter $197X_1$- $197X_2$	Fall $197X_2$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Archival data	W	Decisions must be made about how to record data. For archival information, technical/procedural decisions must be made about precisely what information to obtain (particularly regarding information that will vary from school to school, e.g., parent involvement) and how to code/record	Winter $197X_1$-$197X_2$	End of school year, $197X_2$-$197X_3$
	Structured observations during individual learning tasks (Bronson)	W, (R_{rs})	Some work. Target behaviors chosen from Ogilvie and Shapiro basically for child-child and child-teacher interaction and from Bronson basically for executive skills. Executive skill categories need most work. Decisions must be made about exactly how to sequence observations given varying numbers of subjects per class and per school. Good self-coding score sheets must be devised	Winter $197X_1$-$197X_2$	Fall $197X_2$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Archival data	W	Decisions about what to record and in what form. For archival information, technical/procedural decisions must be made about precisely what information to obtain (particularly regarding information that will vary from school to school, e.g., parent involvement) and how to code/record	Winter $197X_1$-$197X_2$	End of school year, $197X_2$-$197X_3$
	Mastery task (Bronson) or dual focus (Block and Block)	W	Work to be done. Have to decide which of two measures to use	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Complex task (Crandall, Weiner)	W	Work to be done including combining and adapting experimental techniques	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Modeling experiment (Portuges and Feshbach; Ross)	RS	Work on adapting experimental techniques	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Concept-switching or other learning task (Zigler)	RS	Some adaptation work	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Unbalanced/unusual designs (Maw and Maw)	RS	Stimulus selection	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Piagetian egocentrism-sociocentrism task	W, (P_{rs})	Selection of task	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Emmerich role-pictures discrimination task	W, (P_{rs})	Development and adaptation	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Scott pictured value-expectation perception task	W, (P_{rs})	Development and adaptation	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	One of three tasks from the Block battery	RS	Selection from existing tasks	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	"What would your teacher do?" or if not feasible "What happens next?" (Spivak and Shure)	W	The preferred alternative requires development from scratch	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Interview	RS	What to ask and how	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$

00368

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
III. Social and personal development (cont'd)	Alligator game and sentence completion	W	Selection of items and procedures	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Children's achievement wishes test	W	Stimulus selection	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
	Interview	RS	What to ask and how	Winter $197X_1$-$197X_2$	$197X_2$-$197X_3$
IV. Treatment variables	Questionnaire	All classrooms	Little work needed--compile a checklist for: 1. Number of days of treatment 2. Number of hours per day 3. Peer group ethnic composition 4. Turnover in peer group	Late winter $197X_1$	Late winter $197X_2$
	Wiekart/Grannis classificatory scheme for curriculum models	All classrooms	Choose set of alternative models for a classification scheme	Late winter $197X_1$	Late winter $197X_2$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
IV. Treatment variables (cont'd)	PLA-Check	All classrooms	Add two categorization schemes to be applied to each observed activity in classroom: 1. Level of control variables 2. Use of culturally specific materials	Late winter $197X_1$	Late winter $197X_2$
	Tizard natural language observation scheme	Special subsample of classrooms; random sample of English-speaking classrooms only	1. Examination of coding scheme for mutual exclusiveness of categories 2. Coordination of observation of speech (Tizard) with observation of activities (PLA-Check)	Late winter $197X_1$	Late winter $197X_2$
V. Control variables	Questionnaire	All control children	1. Specify alternative types of services supplied children under center and informal day care 2. Develop questionnaire for recommended variables	End Head Start year, $197X_1$	End Head Start year, $197X_2$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
VI. Child background variables	Questionnaire	W	Develop questionnaire or recommended variables	Prior to language battery, end HS year, $197X_1$, for treatment and control children	Prior to language pretests, fall $197X_1$, except attendance variables
VII. Family background	Questionnaire	W	1. Categories of local organizations and local resources 2. Questionnaire for recommended variables	Prior to language battery, end HS year, $197X_1$, for treatment and control children	Prior to language pretests, fall $197X_1$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
VIII. Teacher and teacher aide background variables	Questionnaire	Teachers of all HS classrooms in sample	Construct questionnaire for recommended variables	Winter/ spring $197X_1$	$197X_1$- $197X_2$: optional
	Questionnaire	Teachers of all kindergarten/ first grade classrooms in sample	Construct questionnaire for recommended variables	Winter/ spring $197X_1$	$197X_2$- $197X_3$: immediately prior to ordinary collection of outcome data for children
IX. Center characteristics	Questionnaire	All centers in sample	1. Develop: a. sponsorship categories b. categories for center connections with community c. questionnaire for recommended variables	Winter/ spring $197X_1$	Optional: suggest late winter/ early spring $197X_2$

Table 9-1 (continued)

Battery	Measure	Relevant Sample for Full-Scale Evaluation[a]	Required Development[b]	Proposed Date for Pilot Administration	Proposed Date for Full-Scale Administration
X. Kindergarten/first-grade classroom school characteristics	Questionnaire	All classrooms and schools in which treatment and control children are enrolled	Develop questionnaire for recommended variables	Winter $197X_1$ - $197X_2$	Beginning $197X_2$ - $197X_3$

[a] W = entire sample; RS = random subsample, where subsample consists of a proportional stratified sample of the total sample; SpS = special subsample, where the subsample consists of the total, or proportional subsample of the total sample for particular strata; (R_{rs}) = optional repeated measures on a random sample; (P_{rs}) = optional pretest on a random sample of treatment children only.

[b] For more extensive discussion of required development, see Chapter 3 for health/nutrition measures; Chapter 4 for perceptual/motor/cognitive/language measures; Chapter 5 for social-personal development measures; and Chapter 6 for independent variables.

each center in the United States: (1) county of the center, which gives information about the regional and central city or nonmetropolitan properties of the center; (2) ethnic composition of the center; and (3) amount of treatment provided at the center: five days (half-days or full days) per week; other. Information on type of treatment determines entry of a center onto the sampling list. Information on address and ethnic composition determines the stratum membership of each center on the list. None of this information is available for all centers at the OCD as of July 1974. It will have to be collected from the regional and program offices.

SELECTION OF THE SAMPLE OF SITES FOR THE PILOT EVALUATION

Two decisions must be made before the pilot sample is selected: (1) What strata are to be included in the pilot evaluation--i.e., which of the ten strata will be represented--and (2) should probability or purposive sampling be chosen?

Strata To Be Included

The purpose of the pilot test of the evaluation is to "debug" the procedures required for the full-scale evaluation. Strata may duplicate each other in the problems they present for the evaluation--hostility of the community to outsiders, potential inappropriateness of the tests for children, geographical isolation of sites, etc. Clearly, there is no need to sample from strata that--for purposes of the pilot test--duplicate each other. However, there is no obvious *a priori* similarity between strata on problems. Thus, we recommend sampling from each of the ten strata. In addition, it would be wise to have some experience with inter-site variability within a stratum, simply as a way of assessing the adequacy of the independent variables. For example, we could sample five sites for one stratum of probable intermediate within-stratum variability on the independent variables. The Black central city stratum may be an appropriate choice. In sum, we recommend sampling one site from nine of the strata and five sites from the tenth, where the tenth stratum might be the Black central city stratum. This plan yields a total of 14 sites for the pilot evaluation.

Choice Between Probability and Purposive Sampling of Sites

Random sampling of sites obviously meets the requirement of statistical tests for randomly and independently drawn samples. However, the integrity of statistical tests is not particularly at issue in the pilot test. Both the purpose of the pilot evaluation and the small sample size per stratum argue for centers that provide the maximum information about probable field problems associated with a stratum. What then constitutes maximum information, the worst case or the most probable case? Whatever decision is made about this question, a second question occurs: Which sampling strategy--probability or purposive--is most likely to ensure the selection of centers that yield "maximum information"? Purposive sampling assumes that the researcher knows what he is looking for and how to identify it when he sees it. Random sampling is preferable if there is no agreement about what a maximum information center would be (e.g., worst versus average) or if there is minimal ability to identify such a center.

Our recommendation is to (1) define "maximum information" as "typical," (2) sample the centers randomly from the strata, and (3) check the selection with relevant regional and program offices. Is the selection an outlier? If so, is there reason to think that the deviation seriously jeopardizes the purpose of the pilot test? Where a risk is involved, we suggest that another center be randomly selected from the stratum and the checking procedure repeated.

Pilot evaluation sites are often chosen for their physical proximity to the contractor. This strategy is convenient and usually less expensive for the contractor, advantages that should not be underestimated in a large, complex, and costly field operation. However, the ultimate sample will be geographically scattered. To the extent that field problems are specific to geographical locations, a geographic constraint on the pilot sample reduces what can be learned. Similarly, the contractor does not have the chance to learn about problems associated with "distance from the central office." Thus, if centers are chosen for physical proximity, the choice should be made with an awareness of reduced learning opportunities.

Once centers have been selected and their staffs have agreed to participate in the pilot evaluation, those centers must be removed from the list that will be used in the full-scale evaluation sampling.

PREPARATORY FIELD OPERATIONS

There are three stages of field operations:

- Phase 1 of the pilot test of the evaluation, winter-spring $197X_0$ - $197X_1$.
- Preparation in summer $197X_1$ for the fall $197X_1$ pretests of the full-scale evaluation.
- Phase 2 of the pilot evaluation, winter $197X_1$ - $197X_2$.

Phase 1 of the pilot test involves all procedures and measures required for the first or Head Start year of the full-scale evaluation. The measures consist of all of the health/nutrition and perceptual-motor/cognitive/language measures and all but one of the independent variables batteries. Phase 2 is a trial run of all procedures and measures required for the post-Head Start year of the full-scale evaluation. For some social-personal measures there are optional pretests on treatment children only at the end of the Head Start year. However, all of the nonoptional social-personal development measures are scheduled for $\text{year}_{\text{Head Start + 1}}$ or $\text{year}_{\text{Head Start + 2}}$.

Phase 1 of the Pilot Evaluation

For Phase 1 we consider procedures to be piloted, measures to be administered, and the time schedule for administering the measures.

Procedures. The usual purpose of piloting in research is to "debug" measures. Certainly this is one purpose of the pilot test of an evaluation. However, when a study involves local communities and their members and multiple sites, the procedures for conducting the data collection become crucial.

Table 9-2 provides a list of procedures and a schedule for administering Phase 1 of the pilot evaluation. The full-scale evaluation will include all of the same procedures.

The care with which each of the first seven of these procedures

Table 9-2

PHASE 1, PILOT TEST OF THE EVALUATION: PROCEDURES

Procedure	Relevant Actors	Calendar
1. Contact center in the sample	Contractor; Center Director	As soon as list of centers is compiled and sampling of centers occurs
2. Build community support for the evaluation	Contractor; Center personnel; Head Start parents	If center is willing to consider participating in the pilot evaluation, as soon as the parties can set up a meeting
3. Arrange for a random assignment strategy that the community perceives as fair	Contractor; Head Start parents; Center personnel	Simultaneously with proprocedure 2
4. Hire and train local site coordinator and site evaluation staff for data collection	Contractor; Site Coordinator; Site Evaluation staff; Clerk	As soon as the center and community agree to participate in the evaluation, hiring and training should begin
5. Elicit center personnel and Head Start parents' decisions on site-specific measures	Contractor; Site Coordinator; Center personnel; Head Start parents	Prior to data collection (might be done simultaneously with procedure 2)
6. Elicit center personnel and Head Start parent weightings for cross-site outcomes and criteria for Head Start success	Contractor; Site Coordinator; Center personnel; Head Start parents	Prior to data collection (might be done simultaneously with procedure 2)
7. Keep accurate records of costs	Contractor; Site Coordinator; Clerk	Throughout Phase 1
8. Collect data	Contractor; Site Coordinator; Clerk; Site Evaluation staff; Individuals and groups listed in Column 4, Table 9-3	See Table 9-4

is conducted, especially procedures 2, 3, and 4, strongly affects community acceptance of and the rigor of the eighth procedure, data collection. The remainder of this section discusses these seven procedures in more detail.

Community support. We cannot specify exactly what the contractor should do to build community support for the evaluation because this step is site-specific. However, we can comment on some general guidelines. The doctrine of informed consent that is expected to obtain in medical, psychological, and sociological experiments with human subjects is relevant here. The Head Start evaluation is somewhat different from the usual experiment in that the treatment is a natural event, and informed consent can be assumed from parental request for Head Start. However, the *measurement* of treatment effects is not a usual part of Head Start. Although it might be argued that centers have some obligation to cooperate in the evaluation, the contractor has an obligation to ensure that no person and no group is hurt in the evaluation process. OCD and the contractor will obtain community support for the evaluation only if members of the community trust that they will be protected from biological, psychological, and social harm.

Trust is built in a number of ways. It is important that parents and Head Start personnel know, for example, who is conducting the evaluation, why it is being done, what will be done, what will happen to the results. Specifically, the contractor should be prepared to answer the following standard questions honestly and *in a language that is understandable to all concerned parties:*[1]

- Why is the evaluation being conducted--i.e., who initiated it, what results do they expect, and what do they intend to do with them?
- Who is conducting the evaluation and why--i.e., who is the

---

[1] The contractor might consider producing films for different steps in the process to show members of the local community. For example, there could be a film showing the administration of a measure to a child or test group by a local tester. If films are used, they must be realistic portrayals of what will actually happen.

contractor, how does he intend to benefit from the experiment, and how independent is he of other parties to the evaluation?

- What effect will random assignment have on a child's chances to participate in Head Start?
- What tests will be given to the children and why were they selected?
- What has been done to ensure the cultural fairness of each test?
- What are the maximum number of days and hours per day the child will be tested? Within what time period?
- Where will the child be tested?
- Who will conduct the tests?
- How will the anonymity of the child's data be protected?
- How will the anonymity of the center data be protected?
- What opportunity will the parents of a child have to learn about the child's strengths and medical and other problems, as indicated by the child's test results?
- What opportunity will the center have to learn of the effect of its program?
- How will the data be analyzed for public report--e.g., aggregated by child and by center?

The context in which this information is conveyed is important. Parents and Head Start personnel must have an opportunity to test the contractor's honesty--questions should be encouraged and answered simply and completely; trusted community members who are involved with Head Start should be present to voice group concerns.

One final comment on building community support: Relationships among ethnic groups in the United States are in the process of being redefined. At points in this process, certain groups may show a hostility, anger, and distrust toward outsiders that no amount of contractor sensitivity or openness can counteract. In these cases, the only solution is to eliminate the site from the sample. If the contractor can counteract distrust only by jeopardizing the rights of other constituents--e.g., by jeopardizing the scientific integrity of the evaluation--again the only solution is to eliminate the site from the sample.

Random assignment. Random assignment has to occur prior to the child's entrance into Head Start--i.e., in the summer--and it is this calendar requirement that determines the calendar for procedures 1 through 3. Chapter 7 discusses the problems of random assignment and suggests solutions to these problems. The pilot test of the evaluation provides an excellent opportunity to discuss random assignment with center personnel and Head Start parents. The contractor can note the problems that local constituents perceive, clarify the costs of not assigning children randomly, and explore ways to conduct random assignment that are perceived as equitable, maximize local autonomy, and minimize the administrative burden for the Head Start staff.

Hiring and training for data collection. The quality of data is a function of the reliability and validity of tests for different ethnic groups in natural field conditions and of three field procedures: management of the testing, training of the testers, and monitoring of the testers. We recommend that the contractor hire a site coordinator and clerk from the community to oversee the logistic, or managerial, aspects of testing. The testing will not be satisfactory under the following conditions: (1) the facilities are inadequate (e.g., not enough space, too little privacy, distracting noise or events), (2) the sequence of testing is inefficient (e.g., data on child and family background variables for control children are not scheduled to be collected when parents bring control children in for other test batteries), or (3) the movement of children from the Head Start classroom to the testing room and back or from home and back is not carefully planned (e.g., teachers are not consulted about points in the Head Start day when children can move smoothly in and out of the classroom; explicit transportation arrangements are not made with parents of control children; parents of control children are not called to remind them of their appointment time). Testers might be well qualified, but if they have to operate within a context of perpetual chaos, the quality of data inevitably suffers. There will be missing data and errors in recording responses; administration times per test for children will be uneven; instruction to the child will be less standardized, etc. Smooth testing operations are the result of painstaking and tedious attention to detail. *There is no shortcut.* One criterion for hiring a site coordinator is that person's demonstrated attentiveness to detail.

The second field-specific aspect of data quality is the quality of the testers. First, we strongly recommend hiring testers from the local community. In addition to the economic benefits for the community, well-trained local testers are more likely to procure valid data. We are not concerned with the child's ability to respond to a culturally strange person or to a different dialect, and such factors are undesirable artifacts of the test situation. If the tester is a member of the child's community, the child is apt to feel more comfortable with that person. Second, we recommend a thorough and uniform training of testers, conducted by the contractor's staff and the site coordinator.[1] It is important that the tester understand *how* to administer and record responses to each test, understand the *reasons* for the instructions and specific test items, and appreciate the *necessity* for standardized procedures. Successful training requires careful attention to detail. A test operation can be considered successful when the tester, without help from the training team, can conduct the test and record responses *exactly, not approximately*. The tester must not deviate from the established operations. For example, the tester should not attempt to "help" the shy child by giving him or her "clues" when the test precludes any sort of prompting.

A third field-specific aspect of data quality is systematic monitoring of testers during the testing period. The two reasons for monitoring are well known. First, as the test period progresses, there is a tendency for even conscientious and well-trained testers to become less rigorous. Over time, testers become familiar with the administration of the test and tend to introduce changes in procedure to break the monotony. Second, they may run into situations unanticipated in training. In these cases, testers tend to devise idiosyncratic responses. Obviously, testers cannot and should not be constantly monitored. However, the site coordinator should arrange to observe each tester unobtrusively and periodically throughout the test period.

---

[1] The contractor must assure that training is uniform not only within a site but across sites.

Site-specific measures. We have recommended that center personnel and Head Start parents be encouraged to select one or two measures to be administered in their site only. We saw this option as a way to make the evaluation specifically useful to the local community. We do not know if the current batteries omit child outcomes of great interest to local communities and, if so, what those outcomes might be. The pilot test of the evaluation provides an opportunity to elicit reaction to the idea and to identify particular outcomes of local interest.

Weighting of cross-site outcomes. Several groups of constituents are involved in an evaluation of a national program. They may have different priorities among outcomes--e.g., good health is more important than language development. They may also have different criteria for the success of Head Start with regard to the same outcome--e.g., in the health case, one may want to know if Head Start children are healthier than control children; another, whether the medical problems of Head Start children are diagnosed and, if possible, remedied; a third, whether Head Start children are as healthy as middle-class children.

We have recommended that center personnel and Head Start parents assign their own priorities among the cross-site outcomes and state their criteria for success of outcomes or of particular sets of outcomes. Success criteria may vary depending on whether the outcome is a health or a school-readiness outcome, for example. *Local* program personnel and *poor* communities certainly have priorities among and criteria for the success of local services for children of low-income parents. However, these are not well known at the federal level and among scientific researchers. This is an important chance for these priorities and criteria to surface.

We have several suggestions for eliciting these decisions. Center personnel and Head Start parents should:

- Make their decisions independently. The two groups have different interests and should not be expected to reach consensus on priorities or success criteria.
- Be asked to assign priorities to *blocks* of variables, e.g., health/nutrition outcomes; perceptual-motor/cognitive/language outcomes; social-personal outcomes. The more priorities individuals are asked to assign, the more difficult the task becomes and the more unstable the rankings.
- Be allowed to state criteria for Head Start success without regard to the conventionality of the criterion or the statistical problems involved.

Cost records. The pilot evaluation provides an excellent opportunity to determine cost items and amounts. It is essential that detailed accounts be kept at each site. These records will then allow the contractor and OCD to make final decisions on the test battery and sample size for the full-scale evaluation.

Data collection. In the pilot test the data collection should be structured to estimate field reliabilities and validities of the tests and the success of the training and administrative procedures. Reliabilities can be estimated by inter-judge agreement. Validities can be checked in two ways. First, certain Phase 1 outcomes should be theoretically related to each other, and the scores on indicators of these outcomes should be correlated. Second, testers can observe whether a particular instrument seems to be "picking up" its intended outcome. Staff meetings should be regularly scheduled during the data collection to elicit tester impressions.

The success of training can be estimated in at least two ways. First, the site coordinator and a member of the contractor's staff should monitor testers unobtrusively. Second, for tests with known characteristics, Phase 1 results should be checked for deviations from expected patterns.

Measures. The measures scheduled to be administered during the Head Start year, or Phase 1 of the pilot evaluation, are listed in Table 9-3.

Calendar. Table 9-4 presents the data collection calendar for the independent and dependent variables in Phase 1.

Preparation for the Full-Scale Evaluation

This phase involves preparations for the administration of pretests for the full-scale evaluation in fall $197X_1$. Below we list the steps that have to be taken between the end of the Head Start year, spring $197X_1$, and the beginning of the next Head Start year, fall $197X_1$.

1. On the basis of the experience with the Phase 1 pilot test, make final decisions about which tests to retain in the final battery.
   - o If the measurement battery is too large to be well administered, measure fewer outcomes. One basis for choice is to drop measures with borderline reliability or validity. Another is to examine the intercorrelation matrix for the set of measures and to drop highly redundant ones.
   - o If batteries are too expensive to be administered in the scheduled number of sites, the solutions are to drop either measures or sites.
   - o If measures are not successful in the field, they must be dropped. Measures of known low reliability or questionable validity *should not* be replaced with ones of unknown reliability and validity in order to "salvage" outcomes.
2. Analyze classroom observational data for variation between classrooms and relationships between classroom variation and children's outcomes. If classrooms do not vary on observed variables or if variation seems unrelated to variation in child outcomes, omit classroom observation from the full-scale evaluation.

Table 9-3

PHASE 1, PILOT EVALUATION: MEASURES

Battery	Measure	Sample[a]	Data Source	Measure Administrator
I. Health/nutrition	Immunization records	W	HS medical records for treatment children Parental recall for control children	Evaluation staff
	TB skin test	SpS Native American children only	Child	Medical technician
	Physical examination (optional)	RS	Child	Medical doctor
	Hematocrit	W	Child	Medical technician
	Serum albumin	SpS Native American children only	Child	Medical technician
	Snellen test for visual acuity	W	Child	Nurse
	Pure-tone audiometric screening test	W	Child	Audiometrically trained nurse
	Height, weight, skinfold thickness	W	Child	Nurse

Table 9-3 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator
I. Health/nutrition (cont'd)	Dental examination	W	Child	Dentist
	Treadmill or step test	RS	Child	Medical doctor
II. Perceptual-motor/ cognitive/language	CIRCUS No. 4: Copy What You See	W	Child	Evaluation staff
	CIRCUS No. 3: Look-Alikes	W	Child	Evaluation staff
	CIRCUS No. 13: Thinking It Through	W	Child	Evaluation staff
	CIRCUS No. 2: How Much and How Many	W	Child	Evaluation staff
	CIRCUS No. 5: Finding letters and Numbers	W	Child	Evaluation staff
	CIRCUS No. 1/1a: What Words Mean	SpS English-speaking children, SpS Spanish-speaking children	Child	Evaluation staff

Table 9-3 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator
II. Perceptual/motor/cognitive/language (cont'd)	CIRCUS No. 9/9a: Listen to the Story	SpS English-speaking children SpS Spanish-speaking children	Child	Evaluation staff
	CIRCUS No. 10 (1a, 2): Say and Tell	SpS English-speaking children SpS Spanish-speaking children	Child	Evaluation staff
	CIRCUS No. 10/10a (1b,3): Say and Tell	SpS English-speaking children SpS Spanish-speaking children		
	CIRCUS No. 16: Behavior Inventory	W		Evaluation staff
	Ravens colored progressive matrices	W	Child	Evaluation staff
	Two-person communication game	SpS English-speaking children SpS Spanish-speaking children	Child	Evaluation staff
III. Social and personal development		None administered in Head Start year		

Table 9-3 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator
IV. Treatment variables	PLA-Check	All Head Start classes	Classroom	Classroom observer
	CIRCUS No. 17: Educational Environment Questionnaire	All Head Start classes	Teacher	Evaluation staff
	Tizard natural language observation technique	SpS: subsample of English-speaking Head Start classrooms	Classroom	Classroom observer
	Checklist for properties of treatment, including Weikart/Grannis classification of curriculum models	All Head Start classrooms	Archival information, teacher; observer; center administrator	Classroom observer
V. Control	Questionnaire	All control children	Mother of control child	Evaluation staff
VI. Child background variables	Questionnaire	W	Head Start children; Head Start records; teacher; attendance record Control children parents	Evaluation staff

Table 9-3 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator
VII. Child's family background	Questionnaire	Families of all children	Control children; parents	Evaluation staff
VIII. Teacher background variables	Questionnaire	All Head Start teachers	Head Start teacher	Evaluation staff
IX. Center background characteristics	Questionnaire and checklists	All centers	Center personnel; records of center activities	Evaluation staff
X. Site characteristics	Checklist	All centers	List of centers, by strata	Evaluation staff

[a] W = entire sample; RS = random subsample, where the subsample consists of a proportional stratified sample; SpS = special subsample, where the subsample consists of the total, or a proportional, sample of the total sample for particular strata.

Table 9-4

PHASE 1, PILOT EVALUATION: DATA COLLECTION CALENDAR FOR INDEPENDENT AND DEPENDENT VARIABLES

Calendar	Variables
Late winter $197X_1$	Treatment variables, including classroom observation and, for a special sample, the language environment of the classroom
Winter/spring $197X_1$	Child background variables for Head Start children (except attendance, which must be collected at posttest) Family background variables for Head Start children Teacher background characteristics Center background variables Community characteristics
Spring $197X_1$	Health/nutrition variables Perceptual-motor/cognitive/language variables Control variables Child background variables for control children (prior to administration of language battery) Family background variables for control children (prior to administration of language battery)

3. On the basis of cost data from Phase 1 and other criteria for determining sample size (see Chapter 7), make the final decisions on sample sizes per site and per stratum and on number of strata.
4. If it appears that random assignment is possible in only a small number of sites, either restrict the sample of centers for the full-scale evaluation to those that allow random

assignment or use a value-added design for those sites in the sample that preclude random assignment. The cost of the first choice is generalizability of results. The costs of the second are specified in Chapter 7.

5. If it appears that random assignment is possible in the majority of sites, randomly select the total sample from the stratified list of centers.
6. Randomly select proportional subsamples from the sample for each stratum.
7. If a special subsample involves less than the full sample of the special strata, randomly select the desired proportion from each of those strata.
8. Implement procedures 1-6 (see Table 9-2) for all centers in the sample (contact centers; build community support; arrange for random assignment; hire and train local site coordinator, clerk, and testers; elicit site-specific measures from parents and center personnel; and elicit parent and center personnel weightings and success criteria).

Phase 2 of the Pilot Test of the Evaluation

This phase primarily involves the pilot test of the socioemotional battery for treatment and control children in the same instructional setting, either kindergarten or the first grade. For Phase 2 we consider measures to be administered and procedures involved in the pilot evaluation of Phase 2.

Procedures. Table 9-5 outlines the procedures, actors, and calendar for Phase 2.

Procedures 4 (build local support), 5 (hire and train site coordinator and evaluation staff), and 6 (keep careful cost records) for Phase 2 are the same as their counterparts in Phase 1. Phase 1 procedures 3 (arrange random assignment), 5 (elicit center personnel and Head Start parent choices for site-specific measures), and 6 (elicit center personnel and Head Start parent weightings for cross-site outcomes and criteria for Head Start success) are not relevant to Phase 2 of the pilot run. As indicated, the relevant stage for random

Table 9-5

PHASE 2, PILOT EVALUATION: PROCEDURES

Procedures	Relevant Actors	Calendar
1. Locate the kindergarten or elementary schools that serve Head Start catchment areas for samples of the Phase 1 sites	Contractor	Late summer $197X_1$
2. Contact the relevant kindergartens or elementary schools	Contractor; Administrators of the kindergartens/elementary schools	Late summer $197X_1$ (prior to beginning of school year)
3. Locate classrooms in which treatment and control children are enrolled	Contractor; School administrators; Teachers	Immediately after the beginning of the school year
4. Build local support for the evaluation	Contractor; School administrators; Relevant teachers; Parents	Prior to and immediately after beginning of school year
5. Hire and train site coordinator and evaluation staff for data collection	Contractor; Site Coordinator; Clerk; Site Evaluation staff; Teachers	Prior to winter $197X_1$-$197X_2$
6. Keep careful record of costs	Contractor; Site Coordinator; Clerk	Throughout Phase 2
7. Collect data	Contractor; Site Coordinator; Site Evaluation staff; Teachers	Winter $197X_1$-$197X_2$

assignment is prior to the Head Start year only. Decisions on site-specific measures, cross-site priorities, and success criteria for the total battery should be made once prior to pretests in the Head Start year.

As procedures 1-3 are conducted, it may be found that the treatment and control children scatter widely across schools or across classrooms within a school. The first situation can occur if the Head Start catchment area cuts across school districts; the second, if the area is of sufficiently high density to require multiple classrooms in the school. In this situation it is tempting to eliminate classrooms or schools with very few Head Start and control children from the evaluation. This step might be taken occasionally in the pilot evaluation because the integrity of the study design is somewhat less important there. However, the step should not be taken in the full-scale evaluation. For the full-scale evaluation the integrity of the design in the post-Head Start year is already jeopardized by selective mobility of children and their families out of the geographical area. There are ways to use the Head Start year pretest data to estimate whether the post-Head Start year control and treatment groups are different from each other in their distributions of pretest properties. However, attempts to correct for any bias have all the problems associated with statistical attempts to equate noncomparable control and treatment groups in a quasi-experimental design. Although the evaluation staff has no control over bias introduced by geographical mobility of subjects, it should not introduce bias by eliminating classrooms and schools with small numbers of treatment and control children.

For building support for the evaluation in the kindergarten or first-grade year, several comments are relevant. There are new actors involved in the evaluation: school officials, teachers, and parents of children who are not themselves members of the sample. Since special testing requests are not infrequent in schools, school officials and teachers are more apt to have worked out bases for granting permission for or refusal to these requests, and, in cases where tests are permitted, there are ground rules for the ones to be conducted. The contractor is advised to follow the usual channels for requests of this sort. Since some measures of the socioemotional battery are administered to all children in the kindergarten or first-grade classrooms, children other than Head Start and control children are scheduled to be tested. This raises the question of the informed consent of the parents of these children. Schools usually have standard mechanisms for informing parents of requests to administer special tests to the children. The contractor should use these permission routines unless school officials feel that special procedures should be followed.

The other relevant actors are parents of the control and treatment children. It might be assumed that the parents of these children have already concurred in the post-Head Start year testing. However, in the full-scale evaluation we can expect that a year will have elapsed between the initial parental examination of the test batteries and the administration of the socioemotional battery. This implies that it would be appropriate to go over the test procedures for that battery with the Head Start and control parents again just before its administration.

Measures. As indicated earlier, most post-Head Start year measures are socioemotional measures. For Phase 2, Table 9-6 indicates the battery of tests; measures to be administered; the sample (individuals and groups for whom data are to be collected); the data source (individuals or groups from whom the data are collected); and the administrator of the measure. The currently recommended battery is large. Many of the tests are scheduled for test development, and it is not known how many will be considered suitable for inclusion in the pilot test. However, it must be remembered that as the size of a battery increases, financial costs go up, managerial problems increase, and data quality goes down. Under any circumstances the Phase 2 costs, managerial experiences, and data from the different tests should be carefully evaluated prior to making final choices for the socioemotional battery for the full-scale evaluation. However, if the battery is large, extra attention should be paid to problems contributed by size alone. Even if each measure individually is sound, a large number of such measures can jeopardize the quality of data yielded by any one of them.

Table 9-6

PHASE 2, PILOT TEST OF THE EVALUATION: MEASURES

Battery	Measure	Sample[a]	Data Source	Measure Administrator[c]
I. Social and personal development	Sociometric task: peer nominations	W, (R_{rs})	All children in all classrooms (sample and others)[b]	Observers in the classroom
	Picture naming	W, (R_{rs})	All sample children	Observers in the classroom
	Structured observation during free play (Ogilvie and Shapiro)	W, (R_{rs})	All sample children	Observers in the classroom
	Automatic time-sampling with camera	RS	Children in classrooms (sample and others)	Photos developed and coded by ES; cameras in classroom
	Kelly role construct repertory test	W, (P_{rs})	All teachers, about sample children and their classmates	Group administered, by ES
	Classroom Behavior Inventory	W, (R_{rs})	All teachers, about sample children and their classmates	Group administered, by ES
	Behavior rating	W	All teachers, about all sample children	Self-administered

Table 9-6 (continued)

Battery	Measure	Sample^{a}	Data Source	Measure Administrator^{c}
I. Social and personal development (cont'd)	Large item pool for CBI (e.g., the California Child Q Set)	RS	Teachers, about sample children	Group administered, by ES
	Semantic differential rating of picture stimuli	RS	Teachers, about *unknown* children	Group administered, by ES
	Structured observation during an informal indoors task (Ogilvie and Shapiro, U.S. Commission on Civil Rights; Grotberg appendix)	W, (R_{rs})	All sample children	Observers in the classroom
	Kelly role construct repertory test	W, (P_{rs})	All parents, about their own children	Interviewers
	Behavior rating	W	All parents, about their own children	Interviewers
	Archival data	W	Schools' records, about all sample parents	ES
	Semantic differential using picture stimuli	RS	Parents, about *unknown* children	Interviewers

Table 9-6 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator[c]
I. Social and personal development (cont'd)	Classroom behavior inventory	W, (R_{rs})	Observers, about all children after completing observation sessions	Self administered
	Structured observations during individual learning tasks (Bronson)	W, (R_{rs})	All sample children	Observers in the classroom
	Archival data	W	Schools' records, about all students in sample	ES
	Behavior ratings	W	All teachers, about all sample children	Self administered
	Photo naming and best/worst student nominations	W	All teachers, about all sample children and their classmates	Self administered
	Mastery task (Bronson) or dual focus (Block and Block)	W	All sample children--behavioral	Staff administered (separately)
	Complex task (Crandall, Weiner)	W	All sample children--behavioral	Staff administered (separately)

Table 9-6 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator[c]
I. Social and personal development (cont'd)	Modeling experiment (Portuges and Feshbach; Ross)	RS	Sample children--behavioral	Staff administered (separately)
	Concept-switching or other learning task (Zigler)	RS	Sample children--behavioral	Staff administered (separately)
	Unbalanced/unusual designs (Maw and Maw)	RS	Sample children--preference	Staff administered (separately)
	Piagetian egocentrism--sociocentrism task	W, (P_{rs})	All sample children--self report	Staff administered (separately)
	Emmerich role-pictures discrimination task	W, $(P_{rs})^d$	All sample children--self report	Staff administered (separately)
	Scott pictured value--expectation perception task	W, (P_{rs})	All sample children--self report	Staff administered (separately)
	One of three tasks from the Block battery	RS	Sample children--behavioral or self report	Staff administered (separately)
	"Stuck drawer" (Block and Block)	W	All sample children--behavioral	Staff administered (separately)

Table 9-6 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator[c]
I. Social and personal development (cont'd)	"What would your teacher do?" or--if not feasible--"What happens next?" (Spivak and Shure)	W	All sample children--self report	Staff administered (separately)
	The PIPS test (Spivak and Shure)	W	All sample children--self report	Staff administered (separately)
	Interview	RS	Sample children--self report	Interviewers
	Alligator game and sentence completion (PASS, Minuchin et al.)	W	All sample children--self report	Staff administered (separately)
	Children's Achievement Wishes test (Crandall)	W	All sample children--self report	Staff administered (separately)
	Interview	RS	Sample children--self report	Interviewers
	Self-social Constructs tests (Ziller)	W	All sample children--self report	Staff administered (separately)

Table 9-6 (continued)

Battery	Measure	Sample[a]	Data Source	Measure Administrator[c]
II. Kindergarten/ first-grade school, classroom, teacher variables	Questionnaires	All schools with classes having treatment and control children; all classes with treatment and control children; all teachers of classes with treatment and control children	School administrators, teachers	ES

[a] W = entire sample; RS = random subsample, where the subsample consists of a proportional stratified sample of the total stratified sample; SpS = special subsample, where the subsample consists of the total, or a proportional subsample of the total sample for particular strata; (R_{rs}) = optional repeated measures on a random sample; (P_{rs}) = optional pretest on a random sample of treatment children only.

[b] Kindergarten and first-grade classrooms consist of treatment, control, and other children. Some measures are administered to the total class ("all children in all classrooms"). Others are just administered to the treatment and control children ("all sample children").

[c] ES = site evaluation staff.

[d] For the pre-measure: only family role stimuli will be used.

Calendar. The calendar for Phase 2 of the pilot test of the evaluation is somewhat problematic. Many of the socioemotional measures should be piloted in the fall of $197X_1$ - $197X_2$. This means that pretests of the full-scale evaluation and pilot tests of the socioemotional part of the evaluation will be going on simultaneously. It is certainly administratively impossible to run both of these activities at the same time without jeopardizing the quality of the data for both. *We recommend piloting the socioemotional battery in the winter $197X_1$ - $197X_2$ in a subset of sites used for the Head Start year pilot test.* One advantage of this calendar is that the pilot test of the second year of the full-scale evaluation falls between the pretest and posttest data collections for the first year of the full-scale evaluation. A second advantage is that the socioemotional battery requires most of the test development work. If it is piloted on a small scale during the winter of $197X_1$ - $197X_2$, the contractor has more time to develop the tests before the pilot test.

Two problems with the calendar and reduced scale of the Phase 2 pilot run should be recognized. First, some measures are scheduled in the full-scale evaluation for fall administration in order to minimize the contaminating effects of kindergarten or first grade. If decisions have to be made about measures that are affected by the time of administration, a winter administration is less useful. For example it may be desired to select items that maximally discriminate between treatment and control children. If kindergarten or first grade erodes differences between the two groups, it will be more difficult to locate discriminating items at midyear. Another use of the pilot data may be to select maximally reliable items. Reliability varies with age. Reliability decisions may be less valid if they are based on scores of children who are older than those to whom the measures will be administered.

The second problem is the limited experience with the socioemotional measures. In the decision of sample size for the pilot test of the other two batteries we suggested about fourteen sites, perhaps one each from nine strata and five from one stratum. To the extent that relevant problems with the socioemotional battery differ by stratum, testing the battery in only a few sites reduces the probability that all of the major problems of the full-scale evaluation will be detected in the pilot evaluation.

CHAPTER 10

FOCUSED STUDIES

EVALUATION OF EFFECTS OF HEAD START ON THE HANDICAPPED 389
  Incidence of the Handicap in the Target Population 390
  Accuracy of Identification of Types of Handicaps 393
  Outcomes 395
  Recommendation 397

METACOGNITIVE AND METALINGUISTIC LEARNING IN HEAD START 398

HEAD START EFFECTS FOR SPANISH-SPEAKING CHILDREN 399

HEAD START AS A HEALTH CARE DELIVERY SYSTEM 401

MULTIPLE ROLE INTEGRATION 403

CONVERGENT VALIDATION OF SELECTED CONSTRUCTS 403
  Interview 406
  Rating 407
  Videotape Coding 408
  Self-Role and School-Role Congruence 409

Chapter 10

FOCUSED STUDIES

The initial conception of the focused or small-scale studies was to supplement the national evaluation with research inappropriately conducted in the context of a large evaluation. The studies recommended in this chapter should continue to be seen as adjunct to the national evaluation. However, as indicated in Chapter 1, we have serious reservations about proceeding with a national evaluation of the Head Start program. An alternative perspective for these studies is as elements of a research agenda for a connected set of small studies.

We hope that such an agenda would address questions other than those specified here. The exercise of designing a national evaluation for the four research questions (see Chapters 7 or 8) revealed major gaps in theory about child development and classroom process; in cross-cultural measurement; and in knowledge about the "comparative advantages" of random assignment, value-added, and quasi-experimental designs. For example, attempts to define social competence immediately encountered problems because different constituents define it differently, the term has not been formally explicated within the system of child development concepts, and there is fragmentary empirical knowledge about necessary and sufficient conditions for social competence. A successful attack on question 4 (variations in treatment) requires a number of careful, small experiments with variations in curricular elements of the Head Start program and in classroom processes that might occur within a Head Start environment. Research of this sort allows us to disentangle treatment from site effects, thereby letting us distinguish more efficiently the variations in treatment that affect outcomes from those that do not. We then have a basis for classifying treatments within a program of variable treatments. There needs to be a systematic program of developing versions of theoretically important measures for different socioeconomic and ethnic groups of children. A special methodological study that allows us to compare random assignment, value-added, and quasi-experimental designs would be an important contribution to future evaluations of many kinds of educational interventions, including Head Start. Such a study would require: (1) sufficient sites to construct growth curves, and (2) sites that have sufficient numbers of eligible children to enable both random assignment of volunteers to treatment and control groups and construction of a comparison group of non-volunteer eligible children. Such a study would allow us to elucidate the potential contribution of a value-added design (i.e., specify the scope conditions under which it yields interpretable data). It would also allow an estimate of selection bias, or the "volunteer effect."

As adjuncts to a national evaluation, several suggestions for small-scale studies have been noted throughout the preceding chapters. This chapter describes six of them in greater detail and indicates why each might be considered by OCD. Descriptions take into account that researchers themselves should be free to define the exact nature of the study. Several management recommendations for the small-scale studies are as follows:

- The focused studies should be conducted independently of the large-scale evaluation. These studies are apt to be small and in the vanguard of child development theory and measurement. Thus, they require a different management structure than does a multi-site evaluation.
- To avoid contaminating data and overtaxing both children and personnel participating in the basic evaluation, Head Start centers currently involved in the basic evaluation should not be selected as sites for the focused studies. However, there is no apparent reason why centers formerly involved in a large-scale evaluation could not serve as focused-study sites in subsequent years.
- To determine the generalizability of test results, a subset of measures from the basic evaluation should be administered to children selected for focused study. This will help researchers "locate" subjects for the small-scale studies in the outcome space for the national

evaluation. In other words, researchers can analyze the relationship between small-scale study subjects and subjects for the national evaluation.

o The focused studies should follow the basic evaluation fairly closely in time. This minimizes the extent to which the "history" factor has to be taken into account in estimating the generalizability of the focused studies.

Six specific areas identified for small-scale study are discussed in this chapter: evaluation of effect of Head Start on the handicapped; metacognitive and metalinguistic learning in Head Start; Head Start effects for Spanish-speaking children; Head Start as a health care delivery system; multiple role integration; and convergent validation of selected constructs.

EVALUATION OF EFFECTS OF HEAD START ON THE HANDICAPPED

The assessment of the effects of Head Start on handicapped children presents us with a standard evaluation question. Basic alternative designs for assessing effects are stated in Chapter 7. For the methodological reasons stated there, the random assignment design is the preferable design. Handicapped volunteers for the treatment should be matched on type of handicap and then randomly assigned to the treatment and control groups. However, the handicapped have to be sorted from the nonhandicapped, and types of handicaps have to be sorted from each other. The process by which the potential members of the control group are identified introduces ethical and confounding considerations that preclude our using those children as controls.

There appears to be no way to create a control group by random assignment. The value-added design is not possible because there will not be enough children with the same handicap at different ages to construct growth curves. A modified quasi-experimental design becomes the next-best alternative. In this case, handicapped volunteers are assigned to the treatment. Two groups of controls, each matched with the treatment group on type of handicap, are located from among Head Start-eligible children who did not volunteer for Head Start. The first control group is used for pretest on the dependent variables; the second, for posttest on these variables.

This design is constrained by the following variables:

- Incidence of the specific handicap in the target population;
- Accuracy of the identification of types of handicaps;
- Ability of Head Start to affect the handicap and problems created by it; and
- Importance of the handicap for the child's social competence.

Incidence of the Handicap in the Target Population

Head Start is mandated to serve 10 percent handicapped children. If we assume that the types of handicaps represented in this 10 percent are in proportion to their estimated incidence in the national population of children 0-21 years of age, and if we assume that the incidence of types of handicaps for children 4-5 years old is the same as for the total age range, we can expect the incidence of handicaps in the Head Start children as shown in Table 10-1. We can assume that types of handicaps are evenly distributed across centers unless centers deliberately over- or under-sample certain handicaps or unless certain handicaps are nonrandomly distributed across geographical areas. Both of these distribution biases are possible. Kakalik et al. (1973) discuss the phenomenon of "creaming" in programs for the handicapped (i.e., treating the "easy" handicaps in order to meet quotas). It is also possible that handicaps are unevenly distributed geographically. For example, mental retardation may be overrepresented in "closed" areas because of generations of inbreeding. Nevertheless, even if we assume that types of handicaps are not randomly represented in Head Start centers across the nation (i.e., there is "bunching"), there will be a small incidence of any particular handicap in a Head Start center. If we assume no distribution or selection biases, we can expect a range from one mentally retarded child per 0.9 centers to one visually impaired child per 13 centers. (If we consider the multihandicapped, we

Table 10-1

INCIDENCE OF HANDICAPS IN HEAD START CHILDREN

Type of Handicap	Rate of Occurrence (percent)	Expected Frequency in Head Start[a]	Expected Incidence in Centers[b] Incidence of Handicap per Center	Incidence of Centers per Handicap
Visual impairment	2.02	606	0.08	13.2
Partially sighted	1.88	564	0.07	14.2
Blind	0.13	39	0.01	208.3
Hearing impairment	5.13	1539	0.19	5.2
Deaf	0.52	156	0.02	51.3
Hard of hearing	4.60	1380	0.17	5.8
Speech impairment	23.03	6909	0.86	1.2
Crippling or other health impairment	17.54	5262	0.66	1.5
Mental retardation	29.31	8793	1.10	0.9
Emotional disturbance	15.7	4710	0.59	1.7
Learning disability	7.74	2322	0.29	3.5
Multihandicapped	0.52	156	0.02	51.3

[a] The expected frequency of each type of handicap is based on the rate of occurrence of each type relative to other types, times the expected number of handicapped children in Head Start (10 percent of 300,000 children or 30,000 handicapped children).

[b] There are between 8000 and 9000 centers. Each center has an approximate average of three classrooms. The conservative figure of 8000 centers was used for calculating the incidence of handicap per center and incidence of centers per handicap.

can expect one such child per 51 centers.) This incidence does not respect categories within handicaps. For example, we would expect one emotionally disturbed child per 1.7 centers. However, there are many categories of emotional disturbance. For certain dependent variables the variations between categories of emotional disturbance probably matter. The incidence of each category of emotional disturbance per center will be much lower.

The expected frequencies of different handicaps per center imply problems for a standard evaluation of the effects of Head Start on the handicapped. Whatever outcomes we assess for handicapped children, we have to have adequate sample sizes for each category of handicap for treatment and control groups. Assembling the necessary number of observations has these problems. First, we have to match control and treatment subjects on types of handicaps. As noted above, it is probably impossible to create the control group by randomly assigning volunteers for the treatment to treatment and control groups. If it were possible, matching might be done across geographical areas, but children are then confounded with community characteristics, which increases the variance within blocks. *It may also create noncomparability problems.* If volunteers for the treatment are randomly assigned to the treatment and nontreatment conditions, we assume that the distributions of subject properties that are correlated with selecting the treatment are the same for treatment and control groups. This may be an inappropriate assumption if treatment and control subjects are matched on type of handicap across communities. Reasons for selecting the treatment probably vary from area to area. If so, creating treatment and control groups by matching on handicaps across communities produces groups in which reasons for selecting the treatment are differentially represented. If we are going to go to the trouble of random assignment, it is probably advisable to match on handicaps within communities.

If we are unable to use random assignment, we have to try to equate treatment and control groups as much as possible. Matching on community, as well as on type of handicap, is essential.

Whatever the basis for creating the control group, the pool of Head Start-eligible children in a particular geographical area has to be large enough to allow us to match on handicap within a community.

The second problem is that we have to be able to assess outcomes for the requisite number of subjects at reasonable cost. Given the expected incidence of each type of handicap per center, obtaining the necessary number of observations from a simple random sample of centers will be very expensive. The cost of going into an area decreases as the number of observations that can be made there increases. A simple random sample of centers would put measurement teams into a number of areas that would yield few observations.

We recommend that the evaluation be limited to areas that serve some minimum number of children.

In terms of design, the matching and cost problems end up in the same place. If we are to do a standard evaluation of Head Start effects on handicapped children, the expected incidence of types of handicaps per center restricts the evaluation to children from metropolitan areas. In Region II for OCD there are several metropolitan areas, serving from 290 children in Trenton to 6000 in New York City. Probably New York City is the outlier in terms of number of handicapped children served. If we calculate the incidence of each type for New York City, we have a maximum estimate of how many of each type of handicap we can expect among Head Start participants in a metropolitan area:

Handicap	Expected Incidence in New York City Head Start
Visual impairment	12
Partially sighted	11
Blind	1
Hearing impairment	31
Deaf	3
Hard of hearing	28
Speech impairment	138
Crippling impairment	105
Mental retardation	176
Emotional disturbance	94
Learning disability	46
Multihandicapped	3

Accuracy of Identification of Types of Handicaps

A standard evaluation of handicapped children involves outcomes adjusted for or specific to particular handicaps. Consequently, we have to be able to sort handicapped from nonhandicapped children and types of handicaps from each other. The accuracy of the sorts depends on clarity and mutual exclusiveness of the definition of the handicap, validity and reliability of the measures for detecting its presence, and adequacy of the training of personnel who administer the measures.

Vision-impaired and hearing-impaired children can be sorted out with little error. These are also the lowest-incidence handicaps.

Certain classes of learning disability can be accurately sorted by trained personnel. Other classes are not sufficiently defined to result in an accurate sort. Sorting on emotional disturbance results in a great deal of error. Classes of emotional disturbance are unclearly defined, and measures are unreliable. Mental retardation is better defined than certain classes of learning disabilities and emotional disturbance. However, there is considerable conflict about the validity of standard measures of mental retardation. Children with speech impediments can be sorted out with a fair degree of accuracy.[1] Something about the probable accuracy of sorting on crippling and other health impairments is indicated by the fact that Maternal and Child Health programs have about a 500-item categorization for the child's biological intactness. The potential error in sorting on this number of items is evident.

Kakalik et al. state that if Head Start screening is like the screening in other federal programs for the handicapped, it has a great deal of error in it. They are talking about the error in sorting introduced by inadequately trained personnel. They note that the average pediatrician misses about half of the handicaps in children.

The implications of sorting problems for a standard evaluation are as follows: First, children defined as handicapped by Head Start should be independently screened for the evaluation. Second, for certain kinds of handicaps (e.g., "emotionally disturbed") it is not clear that an independent screening will substantially reduce the sorting errors. Third, quality screening is expensive (e.g., $25 for an audiologist and probably more for a psychiatric evaluation). The quasi-experimental design requires three groups of children who are comparable on handicaps. Since we cannot depend on the Head Start identification of handicaps in the treatment group, this means three waves of screening. Two of those waves are more expensive than the third; locating controls involves screening "out" as well as "in." To the extent that it is difficult to predict in advance of screening whether a child has a particular handicap, there will be costly trial and error in locating controls.

---

[1] Kakalik et al. argue that this is a less important handicap to detect. It tends to correct itself as the child ages, sometimes more rapidly without remediation.

Outcomes

The only point to an evaluation is to assess outcomes that matter and that Head Start can be expected to affect in a consistent way. The question is then: What are those outcomes? *Identifying* the handicap is one outcome. This outcome matters only if the handicap would not otherwise have been identified *and* identifying it helps the child in some way.[1]

A second outcome is *remediation* of treatable handicaps. Curing middle ear infections; fitting glasses, hearing aids, or prosthetics; and operating to correct organic problems fall into this category.

A third set of outcomes derives from *mixing handicapped and nonhandicapped peers* in play and task situations. Some experts have advised that it is good for handicapped children to be with nonhandicapped children in play and task settings. This is an empirical question. Mixing handicapped and nonhandicapped children can have at least two effects. First, unless a child is so severely handicapped that he or she cannot function within a nonhandicapped group, the handicapped child will interact primarily with nonhandicapped children. It usually benefits any individual to be accepted by the social groups to which he or she naturally belongs. We can expect Head Start to affect outcomes related to acceptance (e.g., liking the self) if acceptance varies with being handicapped, and mixing handicapped and nonhandicapped children helps both types of children to ignore the handicap as a basis for inclusion or exclusion.

Alternative effects of mixing are based on the fact that people seem to reach conclusions about themselves by comparing themselves with others. By definition, on some dimension a handicapped child is "less than," "worse at," etc. than his nonhandicapped counterpart. Mixing handicapped and nonhandicapped children provides the handicapped child with a number of comparisons where he or she will come off less well.

---

[1] A child may have a particular handicap about which nothing can be done. If identifying the handicap simply serves to stigmatize the child, the identification process has not been particularly helpful. Similarly, whenever identification goes on, there is some degree of error. A group of Chinese children in California were recently "declassified" as mentally retarded when it was discovered that their IQ scores increased dramatically under native language test conditions.

The *effect* of these comparisons is not clear. They may help handicapped children come to terms with their handicaps, or they may discourage the children from achieving what they could have achieved within the constraints of the handicap.

Other outcomes are less clear. What would we expect Head Start to do for an emotionally disturbed child, for example? This depends on the nature of the emotional disturbance and the problems we would expect that class of disturbance to create for the child, the nature of the adaptation the child has already made to the disturbance, and the teacher's skill in handling the special problems presented by the child. This handicap is extreme in the diversity of problems we would expect children to have. It is also extreme in terms of the diversity of inputs Head Start might make. Since there is little consensus about what a disturbed child needs, we might expect comparable children to elicit different strategies from different teachers. This expected diversity of problems and inputs makes it difficult to specify outcomes relevant to the class--or major subclasses--of emotionally disturbed children.

In order to specify outcomes for other classes of handicaps, we need more information about what can be done for particular handicaps and what we can expect Head Start to do. However, *in general*, if Head Start classrooms and community facilities for the handicapped affect outcomes for handicapped children in Head Start, we can expect variation in outcomes within types of handicaps as a function of variation among Head Start classroom inputs, and variation in outcomes as a function of variation among community facilities if the evaluation is not restricted to metropolitan areas. The argument for the first source of variation is as follows. If the 10 percent rule is observed, there will be a maximum of two handicapped children in a Head Start class. If we assume no selection biases, the probability that both will have the same handicap ranges from 0.0004 for visually impaired to 0.09 for mentally retarded. Thus, to the extent that classroom, including teacher, input affects outcomes, we can expect substantial variation in outcomes within a particular handicap. The argument about the second source of variation is as follows. If the evaluation is restricted to metropolitan areas, community facilities will be held fairly constant within types of handicaps.

To the extent that these facilities affect outcomes for handicapped children, this potential source of variance is of less concern.

Recommendation

A standard evaluation of Head Start effects on handicapped children can be done. A random assignment design or value-added design does not seem possible. Thus, the evaluation is restricted to a pretest and posttest design with two control groups. The control groups are comparable to each other but not to the treatment group. Thus, any difference between treatment and control groups can be attributed to effects other than a treatment effect. The evaluation will also probably have to be restricted to areas with a high density of Head Start-eligible children Since the evaluation is useful only to the extent that it evaluates outcomes for children with particular properties and those properties are infrequent and costly to locate, there is considerable expense associated with the evaluation that is independent of the usual pretest and posttest costs. For certain handicaps (particularly emotional disturbance), the nature of the handicap is unclear. It is also unclear what in a Head Start program might be expected to affect it. These obscurities make it difficult to choose outcomes relevant to the handicap that might be affected by Head Start. For all handicaps, each subject in the treatment group is exposed to a different Head Start classroom. To the extent that Head Start inputs make a difference, we can expect considerable variation within type of handicap on those outcomes we can define. The effect of this is to increase the difficulty of detecting any difference there might be between treatment and control handicapped children.

A standard evaluation may cost more than the data will be worth. A solution might be to assess at least identification of handicaps, remediation of handicaps, and socioemotional effects for a handicapped subsample of the national sample. Data on these outcomes will provide information on handicapped children on those dimensions. It will not tell us whether those things are happening "better," "worse," "faster," etc. to handicapped children in Head Start than to handicapped children not in Head Start.

METACOGNITIVE AND METALINGUISTIC LEARNING IN HEAD START

Participants in the Rand panels on cognitive and language effects agreed strongly that it would be valuable to assess whether Head Start affects metacognitive and metalinguistic learning. Metacognitive skills enable the child to develop conscious strategies for using cognitive abilities: skills in knowing how to seek needed information, how in a given situation to select one appropriate problem-solving strategy from a repertoire of various candidate strategies, how to search memory for needed bits of information, how to present what is known in a form comprehensible to those with whom it it important to communicate. Metalinguistic skills are often closely related, involving the conscious use and manipulation of language capabilities for purposes of effective communication or linguistic play. Both kinds of skills have recently received attention in the developmental literature. Effective use of metacognition and metalanguage indicate a highly adaptive self-awareness and ability to use what one knows. If preschool children have such skills or can acquire them, the children will be at a considerable advantage in the classroom and in all dealings with adults and peers.

Because the measurement of metacognitive and metalinguistic phenomena is in its infancy, and because it is not clear what effects if any Head Start may have in this domain, it is probably unwise to propose any large-scale measurement effort involving the entire national impact study sample. But the potential value of a better understanding of Head Start's influence on metacognition and metalanguage is high enough to merit careful exploration in a controlled substudy. The study should be principally an exercise in hypothesis generation and instrument development, aimed at answering the following questions:

1. What aspects of metacognition and metalanguage can be observed in the Head Start classroom, and how are they adaptive for the child, both at present and later in the elementary school?
2. How can such reasoning and language skills be measured in Head Start so that comparisons can be made among groups of children and among Head Start variations?

3. What aspects of metacognition and metalanguage, if any, can and should be *taught* to preschoolers to help with subsequent schooling? How might this teaching be done?

The study should generate, at a minimum, a clearer operational sense of what we mean by metacognition and metalanguage in Head Start, elucidating constituent skills for teachers and researchers. It also should result in a modest battery of instruments to be used one-to-one or in classroom observation to measure metacognitive and metalinguistic relevance to Head Start. These measures should be sensitive to between-child and between-program variations. Finally, and more ambitiously, as a second step the research project might also be called upon to devise a metacognitive and metalinguistic curriculum for the preschool, suitable for subsequent testing in a controlled experimental setting.

HEAD START EFFECTS FOR SPANISH-SPEAKING CHILDREN

Head Start is not the same for Spanish-speaking children as for children whose only language competency is in English. For Spanish-speaking children, or children whose parents speak a mixture of Spanish and English in the home, Head Start not only must introduce children to new concepts and skills but also must introduce them to an entirely new language environment. Rand has received strong recommendations from panelists and consultants for a special substudy of program effects for this subpopulation.

A study of outcomes for Spanish-speaking children is not easy to design. First, there are problems of *defining the relevant target population*. Should we be concerned with all children of Spanish surname, only those whose parents speak Spanish on most occasions in the home, or perhaps only with children whose parents have recently immigrated to the continental United States? The researcher can imagine a spectrum of increasingly stringent definitions of "Spanish-speaking" and must decide what is the wisest or most useful way to define the term. In addition, of course, there are various culturally distinguishable subgroups within the Spanish-speaking population, often coinciding with geographic locations. For instance, California Chicano children cannot be expected to have exactly the same cultural orientation and dialect as Puerto Rican children, or even as other Chicano children in Texas.

A second major problem is *to determine and categorize the range of program goals* of Head Start programs for Spanish-speaking children. Programs for this subpopulation vary greatly in what they set out to accomplish. Some intend that the child emerge equally competent in Spanish and English and that versatility in skills acquired in the program be demonstrated equally in the Spanish cultural setting and the English-speaking setting of the school; these programs are thoroughly bilingual and bicultural. Other programs are interested only in familiarizing the child with the English-speaking world of school and "dominant" culture. Categories of program objectives should be generated that can lead to empirically based dimensionalization of current Head Start goals.

Third, there is the closely related question of *determining whether measurement should be aimed at concept and skill attainment in the child's own language or in Standard English*. This issue is largely dependent on program objectives; some programs expect to measure success only in English language competency while others strive for complete biculturalism in measurement. Choice of instruments and testers may vary greatly depending on differences in program objectives.

These issues should be clarified. A study of the Spanish-speaking Head Start population should address the following specific questions:

1. What is the most useful way to delimit and define Spanish-speaking Head Start children as a population? How should we define discrete subgroups within the population?
2. What is the best typology of program goals and curricula currently found within Head Start for Spanish-speaking children? How do these differ from one another?
3. Which program objectives, within and across programs, are essentially the same for Spanish-speaking children as for any other children? Which are different, uniquely important for Spanish children because of cultural background differences or cultural values?

4. Which objectives, within and among program types, lend themselves to assessment with instruments designed for all cultural groups, such as the instruments in the basic Head Start battery? Which do not, requiring special adaptation of such instruments, special testers or testing situations, or completely different instruments?
5. In light of various program goals and various desired measurement strategies, how should program success be defined? What constitutes good Head Start program prototypes for Spanish-speaking children, and in general how are good prototypes identified?

Researchers conducting the study should begin by preparing a paper answering questions 1-4, after gathering data on current Head Start programs and consulting with experts in the field. They should then select a reasonable number of subgroups and program types (treatments) to compare, selecting or devising instruments to assess the outcomes of each program, ensuring that some instruments are used across all programs to enable outcome comparisons. Results of the study should be presented in a way that helps Head Start directors and the OCD think more clearly about predictable effects of various programs for various subgroups and about which programs on balance are apt to be most effective.

HEAD START AS A HEALTH CARE DELIVERY SYSTEM

It would be interesting to know whether delivery of health care services to children through Head Start is more efficient, reliable, or complete than delivery through some other mechanism, such as community health centers, private pediatricians, or state or federal programs based on Medicaid provisions. Health care a child receives in Head Start could be compared with the care he or she would receive elsewhere if the program did not exist or were replaced by some other program.

One way to make such comparisons is by mounting a study borrowing the "tracer" methodology originally used by David Kessner and his associates (1974) in their study of children's health care delivery in Washington, D.C. Kessner selected several health care problems of rather high incidence among children (called tracers) that were apt to be representative of a larger set of problems in their treatment and therefore reflected fairly accurately the effectiveness of health care delivery in general. The debilities selected for the study were visual and hearing impairments, middle ear infections, and iron deficiency. Kessner and his colleagues drew a random stratified sample of children from two sections of Washington; various subgroups of the children had access to different systems of health care delivery. It was discovered on the basis of independent assessment of the children's health status (individual physical examinations administered as part of the study) that there were no differences among the subgroups of children in frequencies of the tracer problems or in correct diagnosis and treatment of them. The research team was forced to conclude that the various systems of health care delivery available to the different groups of children were equally effective or, as it turned out, ineffective.

Kessner's study involved school-age children; Head Start therefore was not one of the delivery systems considered. It would be of considerable interest to know if the same finding would hold true for Head Start-eligible children, some attending the program and some using another mode of health care delivery. We might hypothesize that for Kessner's tracer health problems, for instance, or another set of tracer problems, Head Start does make a difference and that children attending the program receive substantially better health care than other children. More precisely, the study might be able to answer the following questions:

1. For a set of tracer health problems, does the incidence of these problems differ among Head Start and non-Head Start children in the Head Start-eligible population?
2. Are the problems apt to be better diagnosed and corrected among children attending Head Start programs than among the others in the eligible group?
3. According to the tracer approach, are there variations in the configuration of health care delivery *within* the Head Start program group that seem to make a difference in the adequacy of care? That is, are some Head Start centers sponsoring better methods of health care delivery than others?

Such a study could be conducted in one or two cities where fewer than half of the Head Start-eligible children actually attended the program, or where other programs for young children were available as options to Head Start.

MULTIPLE ROLE INTEGRATION

The major social situations for the young child are home, neighborhood, and school. In this report we have argued that the primary situation Head Start can be expected to affect is school. There has been concern about Head Start effects on the child's neighborhood and family roles, specifically on the child's ability to integrate the expectations others have of him in these different situations. This is clearly an important outcome for the child. However, the state of theory and measurement in multiple role integration necessitates its investigation in a focused study. Specific suggestions for such a study are discussed in Chapter 5. Briefly, these involve describing the social positions occupied by Head Start-eligible children and the exploratory investigation of children judged to be at the extremes of demonstrated ability to integrate multiple roles--those clearly successful and those clearly unsuccessful.

CONVERGENT VALIDATION OF SELECTED CONSTRUCTS

Focused study of four constructs important in the assessment of social competence in the target population is recommended below. These constructs are of considerable interest and yet are not adequately measured by existing evaluative techniques. Available techniques are primarily verbal. Such methods for assessing social phenomena inevitably risk invoking social desirability biasing (a response set more prevalent, or more congruent with established norms, among majority culture children and thus closely linked with cultural bias in instrumentation). They also become confounded with differences in verbal facility, so that scores may involve a component representing subjects' command of productive language processes. To avoid both sorts of method variance, it is desirable to devise more behavioral-experimental measurement procedures. To the extent that measures so developed produce results converging on those yielded by the more verbal instruments recommended in Chapter 5, confidence in the latter assessments would be increased and a significant contribution toward the development of measures of noncognitive characteristics of preschoolers will have been made.

The first construct recommended for this kind of focused study is the response repertoire in a situation of interpersonal conflict. Chapter 5 provides a discussion of the importance of this construct relative to the aims of Head Start. The instrument recommended for measuring this construct in the subject sample as a whole is a verbal response test devised by Spivak and Shure (1973). The test, described above, proposes an imaginary situation and asks the subject to tell what he might do in such a situation; a number of alternatives are verbally elicited.

It is desirable to examine the feasibility of tests that are less verbal and more behavioral, based on the principle of the Spivak and Shure PIPS test. For example, Millett's (1974) apparatus for studying perception of social power might be adapted for such a purpose. Millett has devised a roadway wide enough for only one model car, with a dead-end turnoff near each car's starting position; she uses this apparatus to observe how persons respond to conflict situations given different initial social power conditions as conveyed in the instructional set. In Millett's experimental situation, one vehicle manipulator (either an adult or peer) is an accomplice, and measures reflect the subject's attempts to deal with the accomplice's behavior. Should the accomplice behave repetitively as the antagonist in the Spivak and Shure story sequences, with the situation allowing for a broad range of verbal and nonverbal responses on the part of the subject, a more naturalistic test of the range of responses to interpersonal problems might be provided.

The second construct proposed for focused study is locus of control. This construct, generated by Rotter in experimental learning studies, has emerged as a significant attitudinal correlate of school achievement in numerous studies (e.g, the Coleman Report, 1966). It is also of considerable theoretical interest, potentially capable of linking such diverse outcomes as depressed school achievement among children and community institution alienation among adults in the lower SES strata, seeing both phenomena as effects of perceived powerlessness. Recent research, however, suggests that perceived locus of control is not common across all situations; rather, the same lower status persons who feel powerless in relation to secondary institution machinery feel significantly more powerful in situations involving only personal efficacy (independent of system responsiveness to those efforts). Clearly, it is important to devise a nonverbal measure of locus of control in order to determine when internal or external attributions are being made, along with the influence of those attributions on behavior strategies and outcomes.

Most of the research paradigms on which Rotter's conclusions are based have to do with learning situations where success-expectancy shifts and extinction curves are typical dependent measures (Phares, 1957; James and Rotter, 1958; Holden and Rotter, 1962; Rotter, 1966). In these classical learning situations, a task is presented and instructional sets vary so that subjects believe outcomes are attributable either to skill (internal control) or to luck (external control); success feedback is manipulated. When expectancy is used as a measure (e.g., subjects are asked to place bets on the outcome of the succeeding trial following feedback on the present one), internal control produces expectations based on past outcomes (subjects use feedback to estimate their own skill level and bet accordingly); external control produces "irrational" shifts in expectancy uncorrelated with past outcomes (subjects often increase their bets after a series of unsuccessful outcomes, apparently believing their bad luck streak is bound to end). Likewise, when the rate at which a behavior is extinguished is used as the dependent measure, skill instructions produce rapid extinction, even with a partial reinforcement schedule. Luck instructions produce slow extinction, subjects thinking they might chance to receive another reward even when they have not received a reward over a series of recent trials.

If similar results should characterize the behavior of younger subjects given comparable experimental procedures, then it would be possible to use expectancy and extinction variables as a basis for inferring the operation of skill or luck attributions during a contrived intellectual task. It is therefore suggested that pilot work attempt to replicate Rotter's results in an appropriate subject population. Successful replication would indicate that further investigatory work should be undertaken to develop an experimental paradigm in which either expectancy or extinction served to index internal versus external control assumptions regarding outcomes on a school-like task. Similar dependent measures should also be used to index internal versus external control assumptions regarding personal-efficacy situations (i.e., situations where skill-based outcomes do not depend on the mediation of a grading system or other institutional machinery for success or failure judgments).

The third construct suggested for focused study is school attitude. Because children will spend approximately 30 hours a week in public school, it is intrinsically important that they find that setting at least tolerable and at best positively rewarding. It has further been indicated in Chapter 5 that school attitudes are instrumentally related to both cognitive and socioemotional outcomes. Three methods for approaching school attitude measurement in focused studies are suggested below.

Interview

While the use of an in-depth interview concerning school attitudes was recommended for subsample study in Chapter 5, an elaboration of one of the suggested procedures is recommended here. Cazden's (1966, 1967) reviews of psycholinguistic techniques for eliciting spontaneous speech from young children includes a picture-interview method in which subjects themselves provide the pictures. Using simple cameras, five-year-olds were able to take photographs themselves with very little instruction. After taking a photograph of one standard scene as directed by the examiner, children were allowed to select and photograph a number of scenes on their own. Developed photographs then became stimulus pictures for an interview. It was found that when a child takes the picture himself, and when he can initiate the discussion, he has much more to say. It is suggested that this technique be adapted for exploring school attitudes. Children in the focused study would be asked to photograph and subsequently discuss school scenes. The research staff conducting the study would be expected to devise standardized ways of posing interview questions or using probes so that a variety of effective responses to the school setting could be investigated. A system for coding those responses, or in some fashion deriving dependent variable values, would also have to be developed.

Rating

Attitudes among older subjects are usually measured by having the subject rate an attitude-object or attitude-statement on an ordinal scale. Although children have not often been regarded as capable of performing such rating tasks, this method was actually used to obtain dependent measures in an attitude study with children in the age range of the proposed research.

The classic Byrne-Nelson similarity-liking model (Jones and Gerard, 1967) served as the basis for experimental attitude testing among kindergarteners by Gaynor et al. (1971). The Byrne-Nelson model predicts that liking will, other things equal, be a function of attitudinal similarity. Gaynor corrobrated this model among five-year-old subjects by first ascertaining their attitudes on four items in an intensive interview (each item had a different focus and included a favorite TV show, favorite sport, favorite school activity, a disliked food). Subsequently, the child listened to a tape on which a child of the same age and sex was heard to express attitudes on these same topics (extent of attitudinal agreement was manipulated to a range of similarity in the subjects from zero to 100 percent).

The child's liking of the taped child was then rated by means of an ingenious adaptation of the seven-point scale. The scale values were represented by a series of blocks stacked in heights ranging from one to six blocks tall, with a wide space separating the first three from the last three stacks. The child was told that if he liked the taped child he should stand in front of the taller row; if he did not like the taped child, he should stand in front of the shorter row; and if he did not have any feelings in either direction he should stand in the middle. A middle choice is equated with the value "4" on the Likert scale and the rating is finished. If the subject chooses either the top values (5 through 7) or bottom values (1 through 3), he is given a token to place on one of the stacks of blocks depending on whether he likes or dislikes the taped child a little, a medium amount, or very much.

Gaynor et al. found that attitude similarity and liking were directly related, and the rating method described above exactly replicated paper-and-pencil rating results for an otherwise identical experimental task with older school children as subjects. The same method could be used in the present study to measure attitudes toward school by contriving stimulus tapes in which children (matching the research subjects in age, sex, and ethnicity) expressed academic values; liking of the taped child would then be treated as attitude scores. Pilot work is needed to determine the validity of the suggested method for the present purposes; should it seem feasible as an attitude measure, additional work would be required to select a small set of attitude items capable of discriminating subjects (i.e., yielding a range of liking scores). It might also be worth exploring whether a child could reasonably be asked to rate two or more such stimulus tapes; if so, attitudes toward more than one focal construct could be elicited.

Videotape Coding

Finally, it was suggested by Paul Ekman[1] that children be presented with school scenes, where stimuli could either be live, ongoing situations (control here posing a difficulty) or else filmed representations of typical school activities. While children were engaged in watching such scenes, they would themselves be subjects of observation by a videotape camera. Subsequently, videotaped reactions of children to the school stimuli could be scored to yield indices of school attitudes. Scoring would take into account gestural-postural behavior and other paralinguistic signals of interest and effect.

Ekman is currently working on the coding of videotaped sequences but does not now have a field-ready instrument. Pilot exploration should attempt to use the methodology of nonverbal communication for developing a coding scheme reliably able to detect attitudinal orientation of children toward presented stimuli. The most recent published accounts of similar efforts involving adult subjects are found in Siegman and Pope (1972). Although the studies reported there are interesting and promising, they indicate the state of the art is very far from producing a coding scheme suitable for evaluation purposes. While we recommended that the possibility be explored of coding videotapes of children to look at school attitudes, we regard it as doubtful that any such instrument could be developed for hypothesis-testing purposes during the period of focused study. Such study is regarded as useful for exploring and developing the videotape coding as an attitude measurement method for young school children.

---

[1] Dr. Paul Ekman, University of California Medical School, San Francisco.

Self-Role and School-Role Congruence

The last construct recommended for assessment in a focused study is self-role and school-role congruence. The importance of integrating self-perceptions and school roles in a manner acceptable to the self was underscored in Chapter 5. At the same time, it was emphasized that little is known about how children approach that important problem. Further, a careful literature review turned up no existing methods for investigating awareness of and responses to the self-role and school-role integration difficulty. However, the World Test devised by Block and Block (1973) to elicit a variety of psychodynamic themes could be adapted to elicit self- and school-themes. The World Test involves observation and rating of children's fantasy play in a sandbox fully equipped with toys, including mother and father dolls, many child figures, home furnishings of all sorts, and other toys. The child is allowed 20 minutes of free play, with a nearby examiner inconspicuously observing and scoring the themes that emerge.

For the present purposes, sandbox equipment could be chosen to elicit school themes by including miniature desks, teacher and child figures, playground items, neighborhood surrounding, etc. A scoring system would have to be devised to represent emergence of play themes related to the self- and school-role integration proposed for measurement. Extensive pilot work would be required to develop such an assessment, and should it be successfully piloted, it would still be very time-consuming and costly to administer.

Appendix A

PANEL PARTICIPANTS AND CONSULTANTS IN THE RAND HEAD START PROJECT

HEALTH AND NUTRITION PANEL

OCTOBER 13-14, 1973

PANELISTS

Roslyn Alfin-Slater
Co-author, issue paper
University of California at Los Angeles

James Carter
Meharry Medical College

Samuel Fomon
University of Iowa

Morris Green
Indiana University School of Medicine

D. B. Jelliffe (not present)
Co-author, issue paper
University of California at Los Angeles

Katherine Messenger
Carnegie Council on Children (New Haven)

David Mundel
Harvard University

A. Frederick North, Jr.
Medical Consultant
Chevy Chase, Maryland

Helen Rodriguez
Lincoln Hospital (NYC)

Nathan Smith
University of Washington

OFFICE OF CHILD DEVELOPMENT, DHEW

Raymond Collins
Program Development Innovations Division

Esther Kresh
Research and Evaluation Division

Linda Randolph
Program Development Innovations Division

Saul Rosoff
Acting Director

OFFICE OF CLINICAL SERVICES, DHEW

Mary Egan
Acting Deputy

RAND STAFF

R. Victoria Arana

Sue Berryman Bobrow

Karen Heald

Roger Levien

Senta Raizen

Barbara Williams

MOTOR/PERCEPTUAL DEVELOPMENT PANEL

OCTOBER 8-9, 1973

PANELISTS

Eugene Abravanel
The George Washington University

Susan Carey-Block
Massachusetts Institute of Technology

Rosslyn Gaines
University of California at Los Angeles

Lila Ghent-Braine
Brooklyn College

Marshall Haith
University of Denver

Herbert Pick (not present)
Author, issue paper
University of Minnesota

Peter B. Pufall
Smith College

Rita G. Rudel
Columbia University Medical School

Susan Rydell
Minnesota Metropolitan State College

Philip Salapatek
University of Minnesota

OFFICE OF CHILD DEVELOPMENT, DHEW

Raymond Collins
Program Development Innovations Division

Jenny Klein
Program Development Innovations Division

Esther Kresh
Research and Evaluation Division

Saul Rosoff
Acting Director

RAND STAFF

R. Victoria Arana

Sue Berryman Bobrow

Sister Gail Cabral, IHM
The Catholic University of America

Karen Heald

Roger Levien

Senta Raizen

Barbara Williams

Robert Yin

LANGUAGE DEVELOPMENT PANEL

NOVEMBER 2-3, 1973

PANELISTS

Paul Ammon
University of California
at Berkeley

Elsa Bartlett
Rockefeller Univeristy

Betty H. Bryant
Educational Testing Service

Courtney B. Cazden
Harvard University

Katrina De Hirsch
Presbyterian Medical Center
(NYC)

Susan Ervin-Tripp
University of California
at Berkeley

Helen Featherstone
Author, issue paper
Huron Institute (Newton Corner,
Mass.)

Lila Gleitman
University of Pennsylvania

William S. Hall
Vassar College

Michael Halliday
University of Illinois
at Chicago Circle

Vera P. John-Steiner
University of New Mexico

Robert Krauss
Columbia University

David McNeil
Princeton University

OFFICE OF THE ASSISTANT SECRETARY FOR PLANNING AND EVALUATION, DHEW

William Prosser
Evaluation for Social
Services/Human Development

OFFICE OF CHILD DEVELOPMENT, DHEW

Raymond Collins
Program Development Innovations
Division

Jenny Klein
Program Development Innovations
Division

Esther Kresh
Research and Evaluation Division

RAND STAFF

R. Victoria Arana

Stephen Barro

Sue Berryman Bobrow

Sister Gail Cabral, IHM
The Catholic University of
America

Karen Heald

Roger Levien

Anthony Pascal

Senta Raizen

Joan Ratteray

COGNITIVE DEVELOPMENT PANEL

OCTOBER 17-18, 1973

PANELISTS

Marion Blank
Rutgers Medical School

Garry Bridge
Columbia University
Teachers College

John Butler
Author, issue paper
Harvard Educational Review

Helen B. Douglas
University of Washington

Sylvia Farnham-Diggory
Carnegie-Mellon University

John Flavell
University of Minnesota

Rochelle Gelman
University of California
at Irvine

Edmund Gordon
Columbia University
Teachers College

Samuel Messick
Educational Testing Service

David Mundel
Harvard University

Sandra Scarr-Salapatek
University of Minnesota

Sheldon White
Harvard University

OFFICE OF CHILD DEVELOPMENT, DHEW

Raymond Collins
Program Development Innovations
Division

Jenny Klein
Program Development Innovations
Division

Esther Kresh
Research and Evaluation Division

RAND STAFF

R. Victoria Arana

Sue Berryman Bobrow

Sister Gail Cabral, IHM
The Catholic University of
America

Karen Heald

Roger Levien

William Lucas

John Pincus

Senta Raizen

SPANISH-SURNAMED PROFESSIONALS' PANEL TO CRITIQUE
INTERIM REPORT
JANUARY 23-24, 1974

PANELISTS

Ernest M. Bernal, Jr.
Bilingual Early Elementary Program
(Austin, Texas)

Josué Cruz, Jr. (not present)
The Child Development Association
Consortium (DC)

Gustavo Gonzalez
Center for Applied Linguistics
(Arlington, Virginia)

Arturo Luis Gutiérrez
Office of International and
Bilingual Education
(Austin, Texas)

Mari-Luci Jaramillo
University of New Mexico

Milton N. Silva
Rutgers University
(New Jersey)

OFFICE OF CHILD DEVELOPMENT, DHEW

Soledad Arenas
Program Development and Innovations
Division

Raymond Collins
Program Development and Innovations
Division

Ramón García
Research and Development

Juán Montoya
Career Development and Technical
Assistance Division

RAND STAFF

R. Victoria Arana

Sue Berryman Bobrow

Senta Raizen

Joan Ratteray

BLACK PROFESSIONALS' CRITIQUE OF INTERIM REPORT:

PANEL I

JANUARY 21-22, 1974

PANELISTS

Patricia Allen
Black Child Development Institute (DC)

Harold Freeman, Jr.
Community Research and Service Center
(Cuny, NYC)

Arvern Moore
Head Start Program, ICS, Inc.
(Holly Springs, Mississippi)

OFFICE OF CHILD DEVELOPMENT, DHEW

Raymond Collins
Program Development and Innovations
Division

Bettyann Harvey
Program Development and Innovations
Division

Clennie Murphy
Regional Support Division

James Robinson
Project Head Start

RAND STAFF

R. Victoria Arana

Sue Berryman Bobrow

Roger Levien

Senta Raizen

Joan Ratteray

BLACK PROFESSIONALS' CRITIQUE OF INTERIM REPORT:

PANEL II

MARCH 29-30, 1974

PANELISTS

Vernon Clark
Child Intervention, Technical Assistant
Development System
(Chapel Hill, North Carolina)

Norman Dixon
University of Pittsburgh

Harold Freeman, Jr.
Community Research and Service Center
(Cuny, NYC)

Asa Hilliard
California State University

Maurine McKinley
Black Child Development Institute

Arvern Moore
Head Start Program, ICS, Inc.
(Holly Springs, Mississippi)

Neloweze Powell
Tuskegee Instutute
(Tuskegee, Alabama)

Ruth Thompson
Center for Human Services
(Cleveland, Ohio)

Robert Washington
Case Western Reserve University

Luther Weems
Morehouse College

Geraldine L. Wilson
New York City Regional Head Start
Training Center

Carl O. Word
Baruch College

OFFICE OF CHILD DEVELOPMENT, DHEW

Raymond Collins
Program Development Innovations
Division

Esther Kresh
Research and Evaluation Division

James Robinson
Project Head Start

RAND STAFF

R. Victoria Arana

Sue Berryman Bobrow

Karen Heald

Roger Levien

Senta Raizen

Joan Ratteray

PANEL ON PROCESS VARIABLES

CHICAGO, APRIL 17, 1974

PANELISTS

Joseph C. Grannis
Teachers College (NYC)

Betty M. Hart
University of Kansas

Barak Rosenshine
University of Illinois (Urbana)

Jane Stallings
Stanford Research Institute

Susan Stodolski
University of Chicago

RAND STAFF

Senta Raizen

Joan Ratteray

GENERAL CONSULTANTS

Scarvia B. Anderson
Educational Testing Service

Pierce Barker
Rand Corporation, Santa Monica

Diane Baumrind
University of California - Berkeley

Tora Kay Bikson
University of California (LA)

Jack and Jeanne Block
University of California - Berkeley

Robert Boruch
Northwestern University

John A. Butler
Harvard Educational Review

Steve Carroll
Rand Corporation, Santa Monica

David K. Cohen
Harvard University

Michael Cole
Rockefeller University

Robert Crain
Johns Hopkins University

Lois-ellin Datta
National Institute of Education

Frances Dendy
Educational Testing Service

Paul Ekman
University of California - Berkeley

Robert D. Hess
Stanford University

Gerald S. Lesser
Harvard Univeristy

Carl Morris
Rand Corporation, Santa Monica

Peter Morrison
Rand Corporation, Santa Monica

Andrew Porter
National Institute of Education

William Rogers
Rand Corporation, Santa Monica

Peter Rossi
Johns Hopkins University

Lee Sechrest
Florida State University

Virginia Shipman
Educational Testing Service

Marshall Smith
National Institute of Education

Ralph Strauch
Rand Corporation, Santa Monica

June Tapp
University of Minnesota

Sheldon White
Harvard University

Edward Zigler
Yale University

Appendix B

ABSTRACT OF RECOMMENDATIONS OF BLACK AND SPANISH-SURNAMED PROFESSIONALS' PANELS

INTRODUCTION

This appendix attempts to abstract the ideas, concerns, and recommendations of the Black and Spanish-surnamed professionals about the proposed evaluation of Head Start. These contributions were expressed at panels convened at The Rand Corporation and in response papers by panel particpants.

The full text contains specific information of interest only to the client and its contractor in structuring the evaluation. However, the material abstracted here is of interest to a more general audience of researchers in education. This abstract represents our attempt to reflect the major concerns of the panelists about the difficulties and dangers inherent in evaluation of culturally different groups and the importance of representing the interests of the Head Start population.

We feel strongly that the panelists' views and reservations must be recognized and answered here and in future research efforts. We made every effort to incorporate their recommendations in our proposed design. How successfully that effort was carried out is yet to be determined.

BLACK PROFESSIONALS' PANEL

Appreciating the potentially damaging political nature of evaluation, the Black panel expressed suspicion and anxiety about a nationwide evaluation of a program involving Black people. Distortion or inappropriate interpretation of the findings was one fear. As one panel member noted, "Black children and their families have suffered from the tendency of policymakers to accept these findings prematurely." Another fear was incompetence. Although qualified researchers may be involved in the evaluation of Head Start, their competence is limited unless they are knowledgeable in the processes of education. A third fear was that the political context of the evaluation and the implicit power relationship between assessor and assessed would reinforce minority communities' sense of powerlessness.

For these reasons, the panel wished to assure quality control and thereby the confidence with which findings would be appropriately interpreted. They recommended adoption of several guidelines in research design:

- The panel suggested a concise and parsimonious research design. Within such a design, consideration should be given to significant independent variables that may enhance the explanatory power of the evaluation. These variables might include parent/child ethnicity, region of country, urban/rural residence, program sponsorship, and health care delivery systems available.
- The panel focused on the makeup of the research teams and their responsiveness to local needs. Most important is the involvement of principled research personnel who have demonstrated skill in early childhood educational or psychological research and an appreciation for relevant community variables. To insure sensitivity to minority and community interests, the teams should represent the multi-ethnicity of the Head Start constituency. Monitoring and review of research efforts should be performed by Black professional groups, parents, and Head Start staff. Facilitation of local objectives in testing and feedback of research results to the local community are important components of research responsiveness.
- Concerns of the panel also centered around the measures to be used. They recommended excluding exploratory measures to avoid increasing the likelihood of uninterpretable but misinterpreted findings and adopting in-depth and longitudinal measures to assess real effects.
- The panel addressed the implications of variations among Head Start children for indicators of the criterion variable, "social competence." White, Black, and Spanish-surnamed populations are significantly represented among Head Start programs. Since each community or each minority may have different

values and therefore different ideas about what constitutes competent behavior, investigating a set of social competence variables across sites is inappropriate. Unfortunately, a comprehensive theory of social competence is lacking and so, therefore, is an analytic framework to guide measurement selection. Given the disagreement among Head Start constituency values, the panel took issue with Rand's initial acceptance of all behaviors as relevant to the domain of competence.

- The panel was also concerned with variation in program input. Whatever child outcome measures are chosen, the findings will be confounded with input, or program, variations. The nature of Head Start is lcoal variation to meet local needs. As the panel pointed out, any service delivery program consists of these subsystems: donor (federal government), service delivery (local projects), and recipients (children and parents). Each subsystem has its own value orientation, its own idea of what the goals of Head Start are and how to achieve them. Not only should input variables be taken into account in understanding outcome variation, but they should help define what outcomes to evaluate.
- Because of the noted differences within the Head Start population and program, a national battery of standardized tests may not be valid. The panel felt that most of the tests considered have been standardized on and measure the skills of middle-class urban children. They contend that no instruments could produce fair measures of diverse groups. Present measures do not take into account minority culture styles and values. For instance, in the measurement of receptive vocabulary, standard English is the accepted criterion; Head Start children will be scored lower using that standard when in fact they are at no communicative disadvantage in their own dialect.

In all of the comments by members of the panel is the need to recognize variation among Head Start programs and participants and to adapt the evaluation efforts to understanding and measuring the implications of these variations.

SPANISH-SURNAMED PROFESSIONALS' PANEL

Like the Black professionals' panel, this panel expressed concern about measures of social competence that are valid across culturally different groups. They too feared the potential harm of the proposed evaluation of Head Start because of the obvious white middle-class orientation in the researchers and the measures suggested. Such measures would only lead to findings detrimental to the Spanish-speaking child. Since 15 percent of the Head Start population is Spanish-speaking, the evaluation cannot ignore the unique competence of children coping in two languages and two cultures. An unanswered question in the minds of the panel is whether competence in one setting (e.g., school) is attained at the expense of competence in the other cultural setting (e.g., the family). This question goes to the heart of Head Start's goals: Is the program designed to help minority children assimilate into the majority culture, or is it designed to foster respect for and development in both cultures? Because of the possible threat of alienation from the family and Spanish-speaking community as a result of assimilation, Head Start risks the mental well-being of the child if it does not appreciate the bicultural skills and competencies of the child.

The specific problems addressed by the response papers focus on the uniqueness of the Spanish-speaking child's learning situation and the difficulties he encounters in language development and its measurement in a bilingual society. For example, Mexican-Americans are typically defined as passive, noncompetitive, and present-oriented. Such descriptors often lead to fallacious interpretation of research data (i.e., as explanations of their lower IQ scores). These stereotypes are unsubstantiated by objective research, especially if one recognizes cultural differences within the Spanish-speaking population. In pointing out the fallacy of stereotypes, the panel was not denying differences between Spanish-speaking and Anglo-American children. The panel was pointing out that the appropriate objective for preschool programs should be to build on bilingual and bicultural strengths rather than attempting to erase cultural differences. This definition of the objective has implications for selecting outcomes for Spanish-speaking children in a national evaluation of Head Start.

One panelist reviewed two dozen articles on research on Mexican-American children. He found a confounding of cultural and social-class differences, lack of attention to the cultural background of the child, no conclusive findings, and therefore the need for a coordinated research effort to understand the testing and measurement problems of bilingual subjects. For instance, because of the likely dominance of English among Spanish-speaking children even by the age of four years, the minimum adjustment called for is testing in the child's dominant language (English, English-Spanish, or Spanish).

BIBLIOGRAPHY

Afifi, A. A., and R. M. Elashoff, "Missing Observations in Multivariate Statistics I: Review of the Literature," *Journal of the American Statistical Association*, Vol. 61, September 1966, pp. 595-604.

-----, "Missing Observations in Multivariate Statistics II: Point Estimation in Sample Linear Regression," *Journal of the American Statistical Association*, Vol. 62, March 1967, pp. 10-29.

Allen, C. E., and L. C. Toomey, "Use of the Vineland Social Maturity Scale for Evaluating Progress of Psychotic Children in a Therapeutic Nursery School," in Robert D. Hess and Roberta M. Baer (eds.), *Early Education: Current Theory, Research and Practice*, Aldine Publishing Company, Chicago, 1968.

Anderson, J., and M. Tindall, "The Concept of Home Range: New Data for the Study of Territorial Behavior," *Environmental Design and Research Association*, Vol. 3, 1972.

Anderson, J. E., "The Long-Term Prediction of Children's Adjustment," in Ira Iscoe and Harold Stevenson (eds.), *Personality Development in Children*, University of Texas Press, 1960.

Anderson, S., "Assessment for Personal and Educational Development," *CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten*, proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973.

Anderson, S., and S. Messick, "Social Competency in Young Children," *Developmental Psychology*, Vol. 10, No. 2, 1974, pp. 282-293.

Armor, D. J., "Theta Reliability and Factor Scaling," in H. L. Costner (ed.), *Sociological Methodology, 1973-1974*, Jossey-Bass, San Francisco, 1974.

Armstrong, J. S. and P. Soelberg, "Some Issues in Sociological Measurement," in Herbert L. Costner (ed.), *Sociological Methodology, 1973-1974*, Jossey-Bass, San Francisco, 1974, pp. 1-17.

Ausubel, D., and P. Ausubel, "Ego Development among Segregated Negro Children," in A. Passow (ed.), *Education in Depressed Areas*, Teachers College, Columbia University Bureau of Publications, New York, 1968.

Baer, D., M. Wolf, and T. Risley, "Some Current Dimensions of Applied Behavior Analysis," *Journal of Applied Behavioral Analysis*, Vol. 2, 1969, pp. 119-124.

Bandura, Albert and R. Walters, *Social Learning and Personality Development*, Holt, Rinehart and Winston, New York, 1963.

Banta, T. J., "Tests for the Evaluation of Early Childhood Education: The Cincinnati Autonomy Test Battery (CATB)," in Jerome Hellmuth (ed.), *Cognitive Studies*, Vol. 1, Brunner/Mazel, New York, 1970.

Barker, R. G., and H. F. Wright, "Psychological Ecology and the Problem of Psychosocial Development," *Child Development*, Vol. 20, 1949, pp. 131-143.

Battle, E., and J. Rotter, "Children's Feelings of Personal Control as Related to Social Class and Ethnic Group," *Journal of Personality*, Vol. 31, 1963, pp. 482-490.

Baumrind, Diana, "Current Patterns of Parental Authority," *Developmental Psychology Monograph*, Vol. 1, No. 4, 1971, pp. 1-103.

-----, "The Development of Instrumental Competence through Socialization," in A. Pick (ed.), *Minnesota Symposium on Child Psychology*, Vol 6, Minneapolis, University of Minnesota Press, 1973.

-----, "Effects of Authoritative Parental Control on Child Behavior," *Child Development*, Vol. 37, 1966, pp. 887-907.

-----, "An Exploratory Study of Socialization Effects on Black Children: Some Black-White Comparisons," *Child Development*, Vol. 43, 1972, pp. 261-267.

-----, and A. Black, "Socialization Practices Associated with Dimensions of Competence in Preschool Boys and Girls," *Child Development*, Vol. 38, 1967, pp. 291-327.

Beatty, W. H. (ed.), *Improving Educational Assessment and an Inventory of Measures of Affective Behavior*, Association for Supervision and Curriculum Development, National Education Association, Washington, D.C., 1969.

Becker, W., and R. Krug, "A Circumplex Model for Social Behavior in Children," *Child Development*, Vol. 35, 1964, pp. 311-332.

Berlyne, D. E., *Conflict, Arousal, and Curiosity*, McGraw-Hill Book Company, New York, 1960.

Biddle, B. J., and E. J. Thomas, *Role Theory: Concepts and Research*, John Wiley and Sons, New York, 1966.

Biddle, B. J., P. Twyman, and E. Rankin, *The Concept of Role Conflict*, State University Arts and Sciences Studies, Social Studies Series No. 11, Stillwater, Okla., 1966.

Bijou, S., and Robert F. Peterson, "Methodology for Experimental Studies of Young Children in Natural Settings," *Psychological Record*, Vol. 19, 1969, pp. 177-210.

Bikson, Tora K., "Minority Speech as Objectively Measured and Subjectively Evaluated," *Proceedings of the American Psychological Association*, 1974.

-----, *The Language Conspiracy: Lexical Styles and Social Status*, Ann Arbor, Michigan: Dissertation Microfilms, 1974b.

-----, "Socializing Collective Behavior Styles Among Preschool Black Children: Relationship Between Research Content and Design Methodology," *Proceedings of the American Psychological Association*, 1974c.

Blalock, Hubert M., "Aggregation and Measurement Error," *Social Forces*, 1971, Vol. 50, pp. 151-165.

-----, *Causal Inferences in Nonexperimental Research*, The University of North Carolina Press, Chapel Hill, 1964.

-----, (ed.), *Causal Models in the Social Sciences*, Aldine Publishing Company, Chicago, 1971.

Block, Jack, and Jeanne Block, "Battery of Procedures for a Longitudinal Study of Ego and Cognitive Development in Young Children," National Institute of Mental Health: NIMH Grant No. MA 16080 Report, Washington, D.C., 1972.

-----, "Ego Development and Provenance of Thought," Grant No. MA 16080 Progress Report, National Institute of Mental Health, Washington, D.C., 1973.

-----, and D. Harrington, "Some Misgivings about the Matching Familiar Figures Test as a Measure of Reflection-Impulsivity," National Institute of Mental Health: Grant No. MA 16080 Report, Washington, D.C., 1973.

Bock, R. D., and E. A. Haggard, "The Use of Multivariate Analysis of Variance in Behavioral Research," in D. K. Whitla (ed.), *Handbook of Measurement and Assessment in Behavioral Sciences*, Addison-Wesley, Reading, Massachusetts, 1968.

Boger, R., and S. Knight, "Socio-Emotional Task Force," Final Report, Office of Economic Opportunity, No. CEC-4118, 1969.

Bonney, M. E., and E. L. Nicholson, "Comparative Social Adjustment of Elementary School Pupils with and without Preschool Training," *Child Development*, Vol. 29, 1958, pp. 125-133.

Booz-Allen & Hamilton Management Consultants, *Retrospective Study of Employee Mobility in Head Start Programs*, Prepared for Office of Child Development, U.S. Department of Health, Education, and Welfare, Washington, D.C., May 1973.

Boruch, Robert, "Bibliography: Illustrative Randomized Field Experiments for Program Planning and Evaluation," prepared for the Social Science Research Council under NSF GI-29843 and AID/CM/ta-C-1055, February 1974.

Bower, Eli M., *Early Identification of Emotionally Handicapped Children in School*, Charles C. Thomas, Springfield, Ill., 1960.

Brandt, Richard M., "Observational Methodology for Evaluation of Early Childhood Programs," *Journal of Research and Development in Education*, Vol. 6, No. 3, Spring 1973, pp. 94-109.

-----, *Studying Behavior in Natural Settings*, Holt, Rinehart, and Winston, New York, 1972.

Brim, Orville, "Socialization through the Life Cycle," in Chad Gordon and Kenneth Gergen (eds.), *The Self in Social Interaction*, Vol. 1, John Wiley and Sons, New York, 1968.

Bronfenbrenner, U., "Motivational and Social Components in Compensatory Education Programs: Suggested Principles, Practices, and Research Designs," in E. Grotberg (ed.), *Critical Issues in Research Related to Disadvantaged Children*, Educational Testing Service, Princeton, 1969.

-----, *A Report on Longitudinal Evaluations of Preschool Programs*, Publication No. (OHD)74-25, U.S. Department of Health, Education, and Welfare, Washington, D.C., 1974.

Bronson, M., *Manual for the Executive and Social Skill Profile for Preschool Children*, Harvard Univeristy, Graduate School of Education, Cambridge, 1973.

Brook, R. H., "Quality of Case Assignment," Ph.D. dissertation, Johns Hopkins University, Baltimore, 1972.

Brownlee, K. A., *Statistical Theory and Methodology*, 2d ed., John Wiley and Sons, Inc., New York, 1965.

Bryk, Anthony S., and Herbert I. Weisberg, "A New Approach to Analyzing Quasi-Experimental Data: Value-Added Analysis," The Huron Institute, Cambridge, Massachusetts, July 1974 (unpublished).

Bunker, J. P. et al., *The National Halothane Study*, Washington, D.C., 1969.

Buros, Oscar K. (ed.), *The Sixth Mental Measurements Yearbook*, Gryphon Press, Highland Park, N. J. 1965.

Butler, Annie L., *Current Research in Early Childhood Education: A Compilation and Analysis for Program Planners*, National Education Association Center, Washington, D.C., 1970.

Butler, John A., *Toward a New Cognitive Effects Battery for Project Head Start*, The Rand Corporation, R-1556-HEW, forthcoming.

Campbell, Donald T., "Methods for the Experimenting Society," paper presented in abbreviated and extemporaneous form to the Eastern Psychological Association, Washington, D.C., April 17, 1971, and to American Psychological Association, Washington, D.C., September 5, 1971.

-----, "Reforms as Experiments," *American Psychology*, Vol. 24, No. 4, April 1969, pp. 409-429.

-----, *Time-Series of Annual Same-Grade Testing in the Evaluation of Compensatory Educational Experiments*, 1970 (unpublished).

-----, and Keith N. Clayton, "Avoiding Regression Effects in Panel Studies of Communication Impact," *Students in Public Communications*, No. 3, Summer 1961, pp. 99-118.

Campbell, Donald T., and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in J. Hellmuth (ed.), *Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child*, Brunner/Mazel, New York, 1970.

-----, "Reply to the Replies," in J. Hellmuth (ed.), *Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child*, Brunner/Mazel, New York, 1970.

Campbell, Donald T., and Peter W. Frey, "The Implications of Learning Theory for the Fade-Out of Gains from Compensatory Education," in J. Hellmuth (ed.), *Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child*, Brunner/Mazel, New York, 1970.

Campbell, Donald T., and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Rand McNally and Company, Chicago, 1963.

Campbell, J. H., "The Effects of Different Types of Training of Visual Discrimination, Auditory Discrimination, and Visual-Motor Coordination on Reading Readiness Test Scores of Kindergarten Children," Ph.D. dissertation, Florida State University; Ann Arbor, Michigan, University Microfilms, No. 71-6977, 1971 (abstract).

Carver, R. P., "Two Dimensions of Test: Psychometric and Edurmetric," *American Psychologist*, 29, 1974, pp. 512-518.

Cassel, R., and G. Martin, "Comparing Peer Status Ratings of Elementary Pupils with Their Guidance Data and Learning Efficiency Indices," *Journal of Genetic Psychology*, Vol. 105, 1964, pp. 39-42.

Cazden, C. B., "On Individual Differences in Language Competence and Performance," *Journal of Special Education*, Vol. 1, 1967, pp. 135-150.

-----, "Some Questions for Research in Early Childhood Education," in J. C. Stanley (ed.), *Preschool Programs for the Disadvantaged: Five Experimental Approaches to Early Childhood Education*, The Johns Hopkins Press, Baltimore, 1972.

-----, "Subcultural Differences in Child Language: An Interdisciplinary Review," *Merrill-Palmer Quarterly*, Vol. 12, 1966, pp. 185-219.

Center for Disease Control, *Ten-State Nutrition Survey, 1968-1970, Volume III: Clinical, Anthropometry, Dental*, Publication No. (HSM) 72-8131, Department of Health, Education, and Welfare, Washington, D.C., 1972a.

-----, *Ten-State Nutrition Survey, 1968-1970, Volume IV: Biochemical*, Publication No. (HSM) 72-8131, Department of Health, Education, and Welfare, Washington, D.C., 1972b.

Cicarelli, Victor, "The Relevance of the Regression Artifact Problem to the Westinghouse-Ohio Evaluation of Head Start: A Reply to Campbell and Erlebacher," in J. Hellmuth (ed.), *Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child*, Brunner/Mazel, New York, 1970.

Cicarelli, Victor et al., *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development*, PB 184328, Westinghouse Learning Corporation-Ohio University (distributed by Clearinghouse for Federal Scientific and Technical Information, Springfield, Virginia), June 12, 1969.

CIRCUS: A Comprehensive Program of Assessment Services for Preprimary Children, Educational Testing Service, Princeton, New Jersey, 1974.

Cohen, David, "Social Experiments with Schools: What Has Been Learned?" in Alice M. Rivlin and T. Michael Timpane (eds.), *Planned Variation in Education: Should We Give Up or Try Harder?* Brookings Institution, Washington, D.C. (forthcoming).

Coleman, J. S., et al., *Equality of Educational Opportunity*, Washington, D.C., 1966.

Conlisk, John, and Harold Watts, "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," *Proceedings from Social Statistics Section, American Statistical Association*, 1969, pp. 150-156.

Cooley, W. W., and P. R. Lohnes, *Multivariate Data Analysis*, John Wiley and Sons, New York, 1971.

Cottrell, L. S., "Interpersonal Interaction and the Development of the Self," in David Goslin (ed.), *Handbook of Socialization Theory and Research*, Rand McNally and Company, Chicago, 1969.

Cowen, E. et al., "The Relation of Anxiety in School Children to School Record, Achievement, and Behavioral Measures," *Child Development*, Vol. 36, 1965, pp. 685-695.

Cramer, E. M., "Significance Tests and Tests of Models in Multiple Regression," *American Statistician*, Vol. 26, 1972, pp. 26-30.

Crandall, Virginia et al., "Children's Beliefs of Their Own Control of Reinforcements in Intellectual-Academic Achievement Situations," *Child Development*, Vol. 36, 1965, pp. 91-109.

Dawe, H. C., "A Study of the Effect of an Educational Program upon Language Development and Related Mental Functions in Young Children," *Journal of Experimental Education*, Vol. 11, No. 2, pp. 200-209, December 1942.

Dittman, L. E., and H. Kyle, *Head Start Planned Variations*, case studies, University of Maryland, 1970-71.

Dohrenwend, Bruce P., and Barbara S. Dohrenwend, *Social Status and Psychological Disorder: A Causal Inquiry*, John Wiley and Sons, New York, 1969.

Dopyera, J., "What's Open About Open Education? Some Strategies and Results," paper presented to the meeting of the American Association of Elementary, Kindergarten, Nursery Educators, Washington, D.C., January 22, 1972.

-----, and M. Lay, "Assessing the Program Environments of Head Start and other Preschool Children: A Survey of Procedures," addendum to Final Report to Office of Economic Opportunity, Head Start Evaluation and Research Contract Number OEO 4120, Syracuse University, Syracuse, New York, 1969.

Dowd, D. J., and S. C. West, "An Inventory of Measures of Affective Behavior," in W. H. Beatty (ed.), *Improving Educational Assessment and an Inventory of Measures of Affective Behavior*, Association for Supervision and Curriculum Development, National Education Association, 1969.

Educational Testing Service, ETS-Head Start Longitudinal Study, *Disadvantaged Children and Their First School Experiences*, Princeton, New Jersey, various years, 1963-1973.

-----, "Longitudinal Study of Disadvantaged Shows Educational Needs," *ETS Developments*, Vol. 21, No. 2, spring 1974.

Edwards, A. L., *Experimental Design in Psychological Research*, 3d ed., Holt, Rinehart and Winston, New York, 1968.

Ekman, P., "Differential Communication of Affect by Head and Body Cues," *Journal of Personality and Social Psychology*, Vol. 2, 1965, pp. 726-735.

Ekstrom, R. B., "Problem Solving and Divergent Production," *CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten*, proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973.

Elwood, P. C., and W. E. Waters, "The Vital Distinction," *Nutrition Today*, Summer 1969, pp. 14-19.

Emmerich, W., "Young Children's Discriminations of Parent and Child Roles," *Child Development*, Vol. 30, 1959, pp. 403-419.

-----, "Continuity and Stability in Early Social Development," *Child Development*, Vol. 35, 1964, pp. 311-332.

-----, "Continuity and Stability in Early Social Development: II. Teacher Ratings," *Child Development*, Vol. 37, 1966, pp. 17-27.

-----, "Personality Development and Concepts of Structure," *Child Development*, Vol. 39, 1968, pp. 671-690.

-----, "Disadvantaged Children and Their First School Experiences: Structure and Development of Personal-School Behaviors in Preschool Settings," ETS-Head Start Longitudinal Study, Princeton, New Jersey, November 1971.

-----, "Preschool Personal-Social Behaviors: Relationship with Socio-Economic Status, Cognitive Skills, and Tempo," ETS-Head Start Longitudinal Study, Princeton, New Jersey, 1973.

Evans, John W., and Jeffry Schiller, "How Preoccupation with Possible Regression Artifacts Can Lead to a Faculty Strategy for the Evaluation of Social Action Programs: A Reply to Campbell and Erlebacher," in J. Hellmuth (ed.), *Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child*, Brunner/Mazel, New York, 1970.

Fanshel, S., *Investigation of the Conceptual and Methodological Problems of a Health Status Index*, Final Report, New York State Health Planning Commission, New York, September 1969.

Farrar, D. E., and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *Review Economics and Statistics*, Vol. 49, 1967, pp. 92-107.

Featherstone, Helen, "Assessing Language Development among Head Start Children," issue paper prepared for the Rand Language Development Panel, Washington, D.C., 1973.

Feinberg, Stephen E., "Randomization and Social Affairs: The 1970 Draft Lottery," *Science*, Vol. 171, 1971, pp. 255-261.

Feshbach, N., "Cross Cultural Studies of Teaching Styles in Four-Year-Olds and Their Mothers," Minnesota Symposium on Child Psychology, University of Minnesota Press, Minneapolis, Minnesota, 1973, pp. 87-116.

-----, and K. Roe, "Empathy in Six- and Seven-Year Olds," *Child Development*, Vol. 39, 1968, pp. 133-145.

Finn, J. D., *Multivariance: Univariate and Multivariate Analysis of Variance, Covariance, and Regression: A Fortran IV Program*, State University of New York, Buffalo, 1968.

Fisher, B., "The Social and Emotional Adjustment of Children with Impaired Hearing Attending Ordinary Classes," *British Journal of Educational Psychology*, Vol. 36, 1966, pp. 319-321.

Fisher, R. A., *The Design of Experiments* (5th ed.), Oliver and Boyd, Edinborough, 1949.

Fitts, W. H., *Tennessee Self Concept Scale*, Counselor Recordings and Tests, Nashville, 1964.

French, J., and B. H. Raven, "The Bases of Social Power," in Darwin Cartwright (ed.), *Studies in Social Power*, University of Michigan Press, Ann Arbor, 1959.

Gaynor, C., J. Lamberth, and J. McCullers, "A Developmental Study of Interpersonal Attraction," Department of Psychology, University of Oklahoma, 1971.

Gerard, H. B., and N. Miller, "Factors Contributing to Adjustment and Achievement in Racially Desegregated Public Schools," National Science Foundation, Grant N. 4-444040-22064 Renewal Proposal, 1971.

Glass, G. V., and P. A. Taylor, "Factor Analytic Methodology," *Review Educational Research*, Vol. 36, 1966, pp. 566-587.

Goldberger, Arthur S., *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*, Institute for Research on Poverty, University of Wisconsin, Madison, April 1972.

-----, *Selection Bias in Evaluating Treatment Effects: The Case of Interaction*, Institute for Research on Poverty, University of Wisconsin, Madison, June 1972.

Goodchilds, J., J. Green, and T. K. Bikson, "School Experience and Ethnicity as Reflected in Assorted Adjustment Measures," *Journal of Social Issues*, 1974 (in press).

Gordon, R. A., "Issues in Multiple Regression," *American Journal of Sociology*, Vol. 73, 1968, pp. 592-616.

Goslin, David (ed.), *Handbook of Socialization Theory and Research*, Rand McNally and Company, Chicago, 1969.

Grannis, J. C., "Autonomy in Learning: An Exploration of Pupil's and Teacher's Role in Different Classroom Environments to Develop Criteria and Procedures for Evaluation in Project Follow-Through," excerpt from final progress Report, Teachers College, Columbia University, New York, February 1973.

Green, Bert F., Jr., and John W. Tukey, "Complex Analyses of Variance: General Problems," *Psychometrika*, Vol. 25, No. 2, June 1960, pp. 127-152.

Grotberg, E. H. (ed.), *Critical Issues in Research Related to Disadvantaged Children*, Educational Testing Service, Princeton, 1969a.

Grotberg, E. H., *Review of Research, 1965 to 1969*, ERIC No. ED-028-308, Project Head Start, Office of Economic Opportunity, Pamphlet 6108-13. Washington, D.C., 1969b.

Guilford, J. P., *Psychometric Methods*, 2d ed., McGraw-Hill Book Company, New York, 1954.

Gurin, P. et al., "Internal-External Control in the Motivational Dynamics of Negro Youth," *Journal of Social Issues*, Vol. 25, 1969, pp. 29-54.

Guttentag, Marcia, "Subjectivity and its Use in Evaluation Research," *Evaluation*, Vol. 1, No. 2, 1973, pp. 60-65.

Handler, E. O., *Preschools and Their Graduates*, Ph.D. dissertation, University of Illinois, 1970.

Hannan, Michael T., *Aggregation and Disaggregation in Sociology*, Lexington Books, D. C. Heath and Co., Lexington, Massachusetts, 1971.

Hauser, Stuart, *Black and White Identity Formation*, John Wiley and Sons, New York, 1971.

Healy, A., "The Influence of Nutritional and Educational Programs on Hemoglobin Values in a Head Start Population over a Five Year Period," paper presented at the 13th Annual Ambulatory Pediatric Association Meeting, San Francisco, May 16, 1973.

Hempel, Carl G., "Fundamentals of Concept Formation on Empirical Science," Vol. II, No. 7, *International Encyclopedia of Unified Science*, University of Chicago Press, 1952.

Henderson, E., B. Long, and R. Ziller, "Self-Social Constructs of Achieving and Nonachieving Readers," *Reading Teacher*, Vol. 19, 1965, pp. 114-118.

Heussenstamm, F. K., and R. Hoepfner, Interperson Perception Test, Form AA, 1969.

Hilton, Thomas L., and Cathleen Patrick, "Cross-Sectional Versus Longitudinal Data: An Empirical Comparison of Mean Differences in Academic Growth," *Journal of Educational Measurement*, Vol. 7, No. 1, Spring 1970, pp. 15-24.

Hoel, Paul G., *Introduction to Mathematical Statistics*, 3d ed., John Wiley and Sons, Inc., New York, 1962.

Holden, K., and J. Rotter, "A Nonverbal Measure of Extinction in Skill and Chance Situations," *Journal of Experimental Psychology*, Vol. 63, 1962, pp. 519-520.

Howell, D., "Significance of Iron Deficiencies: Consequences of Mild Deficiency in Children," in *Extent and Meanings of Iron Deficiency in the U.S.: Summary Proceedings of a Workshop*, Food and Nutrition Board, National Academy of Sciences, Washington, D.C., March 8-9, 1971.

Inkeles, A., "Social Structure and the Socialization of Competence," *Harvard Education Review*, Vol. 36, 1966, pp. 265-283.

-----, "Society, Social Structure and Childhood Socialization" in John A. Clauson (ed.), *Socialization and Society*, Little, Brown and Company, Boston, 1969.

James, W., and J. Rotter, "Partial and 100 Percent Reinforcement Under Chance and Skill Conditions," *Journal of Experimental Psychology*, Vol. 55, 1958, pp. 397-403.

Jensen, J., and L. Kohlberg, *Report of a Research and Demonstration Project for Culturally Disadvantaged Children in the Ancona Montessori School*, Report No. 1284, Office of Economic Opportunity, Washington, D.C., 1966.

Jones E., and H. Gerard, *Foundations of Social Psychology*, John Wiley and Sons, New York, 1967.

Jungeblut, Ann, "Quantitative and Relational Understanding; Perceptual Skills," *CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten*, proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973.

Kagan, J., "Preschool Enrichment and Learning," *Interchange*, Vol. 2, 1971, pp. 12-22.

Kakalik, J. S. et al., *Services for Handicapped Youth: A Program Overview*, The Rand Corporation, R-1220-HEW, Santa Monica, May 1973.

Katz, I., "Factors Influencing Negro Performance in the Desegregated School," in M. Deutsch et al. (eds.), *Social Class, Race and Psychological Development*, Holt, Rinehart and Winston, New York, 1968.

-----, "Review of Evidence Relating to Effects of Desegregation on the Intellectual Performance of Negroes," *American Psychologist*, Vol. 19, 1964, pp. 381-399.

Kelly, George A., *The Psychology of Personal Constructs*, Norton, New York, 1955.

Kelly, H. H., "Attribution Theory in Social Psychology," in D. Levine (ed.), *Nebraska Symposium on Motivation*, University of Nebraska Press, Lincoln, 1967.

Kessner, David M., and Carolyn E. Kalk, *Contrasts in Health Status, Volume 2: A Strategy for Evaluating Health Services*, Institute of Medicine, National Academy of Sciences, Washington, D.C., 1973.

Kessner, D. M., C. K. Snow, and J. Singer, "Assessment of Medical Care for Children," *Contrasts in Health Status*, Vol. 3, National Academy of Sciences, Institute of Medicine, Washington, D.C., 1974.

Kidd, Alice H., and R. M. Kidd, "The Development of Auditory Perception in Children," in Alice H. Kidd and J. L. Rivoire (eds.), *Perceptual Development in Children*, International Universities Press, Inc., New York, 1966.

Kimbrough, J., S. Barge, T. K. Bikson, and E. Smith, "The Children's Collective: Progress Report," Office of Child Development, DHEW, Grant No. 488, 1974.

Kirk, Roger E., *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole Publishing Co., Belmont, California, 1969.

Klitgaard, Robert E., *Going Beyond the Mean in Educational Evaluation*, The Rand Corporation, Santa Monica, P-5184, March 1974.

Kohlberg, L., "Stage and Sequence: The Cognitive-Developmental Approach to Socialization," in David Goslin (ed.), *Handbook of Socialization Theory and Research*, Rand McNally and Company, Chicago, 1969.

Kohlberg, L., J. LaCrosse, and D. Ricks, "The Predictability of Adult Mental Health From Childhood Behavior," in Benjamin B. Wolman (ed.), *Manual of Child Psychopathology*, McGraw-Hill Book Co., New York, 1972.

Kohn, M., and B. Rossman, "A Social Competence Scale and Symptom Checklist for the Preschool Child," *Developmental Psychology*, Vol. 6, 1972, pp. 430-444.

Kreutzer, M. A., Sister C. Leonard, and J. H. Flavell, "An Interview Study of Children's Knowledge about Memory," in J. H. Flavell et al., *Metamemory and Related Information Processing and Metacognitive Phenomena*, Institute for Child Development, University of Minnesota, 1974.

Labov, W., "The Logic of Nonstandard English," in Frederick Williams (ed.), *Language and Poverty*, Markham Press, Chicago, 1970.

Lamb, H. et al., *The Development of Self-Other Relationships During Project Head Start*, ERIC No. ED015008, University of Delaware, Newark, Delaware, 1965.

Lambert, N., "The Development and Validation of a Process for Screening Emotionally Handicapped Children in School," Office of Education, Washington, D.C., Cooperative Research Project No. 1186, 1962.

Leeper, R., "A Study of a Neglected Portion of the Field of Learning: The Development of Sensory Organization," *Journal of Genetic Psychology*, Vol. 46, 1935, pp. 41-75.

Levy, P., "Substantive Significance of Significant Differences between Two Groups," *Psychological Bulletin*, Vol. 67, 1967, pp. 37-40.

Light, Richard J., and Paul V. Smith, "Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies," *Harvard Educational Review*, Vol. 41, No. 4, November 1971, pp. 429-471.

-----, "Choosing a Future: Strategies for Designing and Evaluating New Programs," *Harvard Educational Review*, Vol. 40, No. 1; February 1970, pp. 1-28.

Lindquist, E. F., *Design and Analysis of Experiments in Psychology and Education*, Houghton Mifflin Company, Boston, 1953.

Linn, E. L., *The Socially Disadvantaged Child: Teacher Correlates*, Ph.D. dissertation, University of Texas, Austin, 1966.

Lipton, E. D., "The Effect of a Physical Education Program to Develop Directionality of Movement on Perceptual-Motor Development, Visual Perception, and Reading Readiness of First Grade Children," Ph.D. dissertation, New York University; University Microfilms, Ann Arbor, Michigan, No. 69-21, 1969.

Lord, Frederic M., "Large-Sample Covariance Analysis When the Control Variable Is Fallible," *Journal of the American Statistical Association*, Vol. 55, June 1960, pp. 307-321.

-----, "A Paradox in the Interpretation of Group Comparisons," *Psychological Bulletin*, Vol. 68, No. 5, 1967, pp. 304-305.

-----, "Statistical Adjustments when Comparing Preexisting Groups," *Psychological Bulletin*, Vol. 72, No. 5, 1969, pp. 336-337.

Maccoby, E., "Role-Taking in Childhood and its Consequences for Social Learning," *Child Development*, Vol. 30, 1959, pp. 239-252.

-----, "What Copying Requires," *Ontario Journal of Educational Research*, No. 10, 1968, pp. 163-170.

Madow, William, *Britannica Review*, 1969.

Madsen, M., "Developmental and Cross Cultural Differences in the Cooperative and Competitive Behavior of Young Children," *Journal of Cross Cultural Psychology*, Vol. 2, 1971, pp. 365-371.

Mahoney, F. I., and D. W. Barthel, "Functional Evaluation: The Barthel Index," *Maryland State Medical Journal*, Vol. 14, February 1965, pp. 61-65.

Marascuilo, L. A., and J. R. Levin, "Appropriate Post Hoc Comparisons for Interaction and Heated Hypotheses in Analysis of Variance Designs: The Elimination of Type IV Errors," *American Education Research Journal*, Vol. 7, 1970, pp. 397-421.

Marschak, M., "Teachers Evaluate the Progress of the Head Start Child," Economic Youth Opportunities Association, Washington, D.C., 1966.

Maw, W., and E. Maw, "Selection of Unbalanced and Unusual Designs by Children High in Curiosity," *Journal of Child Development*, Vol. 33, 1962, pp. 917-922.

McCandless, Boyd, "Discussion," *CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten*, proceedings of a symposium presented at the American Psychological Association, Montreal, Quebec, August 21, 1973, pp. 39-40.

McClelland, David C. et al., *The Achievement Motive*, Appleton, Century, Crofts, New York, 1953.

McElroy, Donald L., and William F. Malone, *Handbook of Oral Diagnosis and Treatment Planning*, Williams and Wilkins, Baltimore, 1969.

McNeil, K., and B. Phillips, "Scholastic Nature of Responses to the Environment in Selected Subcultures," *Journal of Educational Psychology*, Vol. 60, 1969, pp. 79-85.

McNeill, David, *The Acquisition of Language*, Harper and Row, New York, 1972.

McNemar, Quinn, "A Critical Examination of the University of Iowa Studies of Environmental Influences Upon the IQ," *Psychological Bulletin*, Vol. 37, No. 2, February 1940, pp. 63-92.

Medley, D. M. and N. E. Mitzell, "Measuring Classroom Behavior by Systematic Observation," in N. Gage (ed.), *Handbook of Research on Teaching*, Rand McNally and Co., Chicago, 1963.

Medley D. M., C. Schluck, and N. Ames, *Recording Individual Pupil Experiences in the Classroom: A Manual for PROSE Recorders*, Educational Testing Service, Princeton, December 1968.

Meehl, Paul E., "Nuisance Variables and the Ex Post Facto Design," in Michael Radner and Stephen Winokur (eds.), *Minnesota Studies in the Philosophy of Science*, Vol. IV, University of Minnesota Press, 1970, pp. 373-402.

Mercer, Jane, "Socio-Cultural Correlates of Learning and Behavior 'Problems,'" prepared for Conference on Learning Disabilities and Behavior Problems, sponsored by the National Institute of Education, Washington, D.C., June 1974.

Messick, Samuel, "The Context of Assessment and the Assessment of Context," *CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten*, proceedings of a symposium presented at the American Psychological Association, Montreal, Quebec, August 21, 1973, pp. 33-38.

Meyer, Donald L., and Raymond O. Collier, Jr. (eds.), *Bayesian Statistics*, F. E. Peacock Publishers, Inc., Itasca, Illinois, 1970.

Meyers, C. E. et al., "Four Ability-Factor Hypotheses," *Monographs of the Society for Research in Child Development*, Vol. 29, Item 5, 1964, Serial No. 96.

Miller, K., and R. M. Dreger, *Comparative Studies of Blacks and Whites in the United States*, Academic Press, New York, 1973.

Miller, L. B., et al., *Experimental Variation of Head Start Curricula: A Comparison of Current Approaches*, Progress Report No. 9, University of Louisville, March-May, 1971.

Millett, J., "Social Power Analysis of the Dyad in the Classroom Setting," *Proceedings of the Western Psychological Association*, 1974.

Minuchin, Patricia et al., *The Psychological Impact of School Experience: A Comparative Study of Nine-Year-Old Children in Contrasting Schools*, Basic Books, New York, 1969.

Moreno, J. L., *Who Shall Survive? A New Approach to the Problems of Human Interrelations*, Nervous and Mental Disease Publishing, Washington, D.C., 1934.

Morrison, D. F., *Multivariate Statistical Methods*, McGraw-Hill, New York, 1967.

Mosteller, F., and R. R. Bush, "Selected Quantitative Techniques," in G. Lindsey (ed.), *Handbook of Social Psychology*, Vol. 1, Addison, Wesley, Cambridge, Massachusetts, 1954.

Mussen, Paul H., *Handbook of Research Methods in Child Development*, John Wiley and Sons, New York, 1960.

National Academy of Sciences, *The Relationship of Nutrition to Brain Development and Behavior*, a position paper of the Food and Nutrition Board, National Academy of Sciences, Washington, D.C., June 1973.

National Center for Health Statistics, *Age Patterns in Medical Care, Illness, and Disability, United States, 1968-1969*, Publication No. (HSM) 72-1026, Series 10, Number 70, Department of Health, Education, and Welfare, Washington, D.C., 1972a.

-----, *Characteristics of Persons with Corrective Lenses, United States, July 1965-June 1966*, Publication No. 1000, Series 10, Number 53, Public Health Service, Washington, D.C., 1969.

-----, *Plan and Initial Program of the Health Examination Survey*, Publication No. 1000, Series 1, Number 4, Public Health Service, Washington, D.C., 1965.

-----, *Quality Control in a National Health Examination Survey*, Publication No. (HSM) 72-1023, Series 2, Number 44, Department of Health, Education, and Welfare, Washington, D.C., 1972b.

-----, *Visual Acuity of Children, United States*, Publication No. 1000, Series 11, Number 101, Public Health Service, Washington, D.C., 1970.

-----, *Volume of Physician Visits, United States, July 1966-June 1967*, Publication No. 1000, Series 10, Number 49, Public Health Service, Washington, D.C., 1968.

Neyman, J., "Basic Ideas and Theory of Testing Statistical Hypotheses," *J. Royal Statistical Society*, 105, 1942, pp. 292-327.

Nunnally, J. G., *Psychometric Theory*, McGraw-Hill, New York, 1967.

Office of Child Development, Department of Health, Education, and Welfare, *OCD-Head Start Policy Manual*, January 1973.

Ogilvie, D., and B. Shapiro, "Manual for Assessing Social Abilities of One-to Six-Year-Old Children," in B. L. White et al. (ed.), *Preschool Project: Child Rearing Practices and the Development of Competence*, Office of Economic Opportunity, Washington, D.C., Final Report, Grant No. CG-9909, 1972.

Osgood, Charles, G. Suci, and P. Tannenbaum, *The Measurement of Meaning*, University of Illinois Press, Urbana, Ill., 1958.

Overall, John E., and Douglas K. Spiegel, "Concerning Least Squares Analysis of Experimental Data," *Psychological Bulletin*, Vol. 72, No. 5, 1969, pp. 311-322.

Pascale, M. A., "The Effect of a Visual-Motor Integration Training Program on Beginning Writing Skills of Kindergarten Children," Ph.D. dissertation, Columbia University; University Microfilms, Ann Arbor, Michigan, No. 70-12, 1970.

Phares, E. J., "Expectancy Changes in Skill and Chance Situations," *Journal of Abnormal and Social Psychology*, Vol. 54, 1957, pp. 339-342.

Piaget, Jean, *The Moral Judgment of the Child*, Free Press, Glencoe, Ill., 1948.

-----, *The Psychology of Intelligence*, Routledge, Kegan Paul, London, 1947.

Piers, E., and D. Harris, "Age and Other Correlates of Self-Concept in Children," *Journal of Educational Psychology*, Vol. 55, 1964, pp. 91-95.

-----, *Manual for the Piers-Harris Children's Self Concept Scale*, Counselor Recordings and Tests, Nashville, 1969.

Polling, H. M., "Auditory Deficiencies in Poor Readers," in Helen M. Robinson (ed.), *Clinical Studies in Reading* (Supplementary Educational Monograph), University of Chicago Press, Chicago, 1968.

Porter, A. C., *The Effects of Using Fallible Variables in the Analysis of Covariance*, University Microfilms, Ann Arbor, Michigan, 1967.

Portuges, S., and N. Feshback, "The Influence of Sex and Socioethnic Factors Upon Imitation of Teachers by Elementary Schoolchildren," *Child Development*, Vol. 43, 1972, pp. 981-989.

Prescott, E. et al., *Assessment of Child-Rearing Environments: An Ecological Approach*, Pasadena, California, Pacific Oaks College, Prepared for Children's Bureau, Office of Child Development, U.S. Department of Health, Education, and Welfare, June 1971.

Proshansky, H., and P. Newton, "The Nature and Meaning of Negro Self-Identity," in M. Deutsch et al. (eds.), *Social Class, Race and Psychological Development*, Holt, Rinehart and Winston, New York, 1968.

Radin, N., and D. P. Weikart, *A Home Teaching Program for Disadvantaged Preschool Children*, Journal of Special Education, Vol. 1, No. 2, pp. 183-195, 1967.

Research Triangle Institute, *Some Aspects of Child Development of Participants in Full Year 1967-68 and 1968-69 Programs*, Final Report, Project No. 22U-581, Durham, North Carolina, December 1972.

Riecken, H. W. et al., *Experimentation as a Method for Planning and Evaluating Social Programs*, Evanston, Ill., Seminar Press for the Social Science Research Council, 1974.

Risley, T., and M. F. Cataldo, "Evaluation of Planned Activities: The Pla-Check Measures of Classroom Participation, in Davidson, Clark, and Hammerlynck (eds.), *Evaluation of Social Programs in Community, Residential and School Settings*, Research Press, Champaign, Illinois, 1974.

-----, *Planned Activity Check*, Center for Applied Behavior Analysis, 1973.

Robinson, J. P., and P. R. Shaver, *Measures of Social Psychological Attitudes*, University of Michigan Institute for Social Research, Ann Arbor, 1974.

Rosenshine, B., *Teaching Behaviors and Student Achievement*, National Foundation for Educational Research in England and Wales, Windsor, Berkshire, England, 1971.

-----, and N. Furst, "The Use of Direct Observation to Study Teaching," in M. W. Robert Travers (ed.), *Second Handbook of Research on Teaching*, Rand McNally, Chicago, 1973.

Rosenthal, R. et al., "Assessing Sensitivity to Nonverbal Communication: The PONS Test," *Newsletter*, American Psychological Association, Division 8, January 1974.

Rosenthal, Robert and Elenore Jacobson, *Pygmalion in the Classroom*, Holt, Rinehart and Winston, New York, 1968.

-----, "Self-Fulfilling Prophecies in the Classroom: Teachers' Expectations as Unintended Determinants of Pupils' Intellectual Competence," in M. Deutsch et al. (eds.), *Social Class, Race, and Psychological Development*, Holt, Rinehart and Winston, New York, 1968.

Rosnow, I., and N. Breslau, "A Guttman Health Scale for the Aged," *Journal of Gerontology*, Vol. 21, October 1966, pp. 556-559.

Ross, D., "Relationship Between Dependency, Intentional Learning, and Incidental Learning in Preschool Children," *Journal of Personality and Social Psychology*, Vol. 4, 1966, pp. 374-381.

Rossi, Peter H., and Walter Williams (eds.), *Evaluating Social Programs*, Seminar Press, New York, 1972.

Rotter, J. B., "Generalized Expectancies for Internal Versus External Control of Reinforcement," *Psychology Monographs*, Vol. 80, 1966, No. 609.

-----, "Some Implications of a Social Learning Theory for the Prediction of Goal-Directed Behavior from Testing Procedures," *Psychological Review*, Vol. 67, 1960, pp. 301-316.

Rozeboom, William W., "The Fallacy of the Null-Hypothesis Significance Test," in Bernhardt Lieberman (ed.), *Contemporary Problems in Statistics*, Oxford University Press, New York, 1971, pp. 116-126.

Rubin, Donald B., "Matching to Remove Bias in Observational Data," *Biometrics*, Vol. 29, March 1973, pp. 159-183.

Ryan, Sally (ed.), *A Report on Longitudinal Evaluations of Preschool Programs: Volume 1, Longitudinal Evaluations*, Publication No. (OHD)74-24, U.S. Department of Health, Education, and Welfare, Washington, D.C., 1974.

Sanders, B. F., "Measuring Community Health Levels," *American Journal of Public Health*, Vol. 54, July 1964, pp. 1063-1070.

Sarason, S. B. et al., *Anxiety in Elementary School Children*, John Wiley and Sons, New York, 1960.

Sarbin, T. R., "Role Theoretical Interpretation of Psychological Change," in P. Worchel and D. Byrne (eds.), *Personality Change*, John Wiley and Sons, New York, 1964.

-----, "Role Theory," in Gardner Lindzey and E. Aronson (eds.), *Handbook of Social Psychology*, Vol. 6, Addison-Wesley, Reading, Mass., 1968.

Schaefer, E. S., "Converging Conceptual Models for Maternal Behavior and for Child Development," in J. C. Glidewell (ed.), *Parent Attitudes and Child Behavior*, Charles C. Thomas, Springfield, Ill., 1961.

Scheffe, H. A., "A Method for Judging All Possible Contrasts in the Analysis of Variance," *Biometrika*, Vol. 40, 1953, pp. 87-104.

Schonell, F. J., and F. E. Schonell, *Backwardness in Basic Subjects*, Oliver and Boyd, London, 1946.

Scott, E., "The Influence of Nursery School Experience in Social Value Acquisition in Preschool Children," *Educational Review*, Vol. 21, 1969, pp. 226-233.

Shallenberger, P., and E. Zigler, "Rigidity, Negative Reaction Tendencies and Cosatiation Effects in Normal and Feebleminded Children," *Journal of Abnormal and Social Psychology*, 1961, pp. 20-26.

Shipman, Virginia, "Disadvantaged Children and Their First School Experiences," Interim Report, ETS-Head Start Longitudinal Study, Princeton, August 1973.

Siegman, A., and B. Pope (eds.), *Studies in Dyadic Communication*, Pergamon Press, New York, 1972.

Silver, Henry K., C. H. Kempe, and H. B. Bruyn, *Handbook of Pediatrics* (Tenth ed.), Lange Medical Publications, Los Altos, California, 1973.

Simon, A. and E. G. Boyer (eds.), *Mirrors for Behavior: An Anthology of Classroom Observation Instruments, Vols. 1-6*, Research for Better Schools, Philadelphia, 1967.

-----, *Mirrors for Behavior: An Anthology of Classroom Observation Instruments, Vols. A and B, 7-14, and Summary*, Research for Better Schools, Philadelphia, 1970.

Smith, D., "The Effect of Selected Communication Patterns on the Level of Abstraction, Length, and Complexity of Sentence and Speech of Children," ED-028-004, Educational Resource Information Center, 1969.

Smith, J. L., *Nutritional Status of New Orleans, Mississippi and Alabama Head Start Children*, Tulane University Head Start Evaluation and Research Center, Final Report to the Office of Economic Opportunity, August 31, 1969.

Smith, M. Brewster, "Competence and Socialization," in J. A. Clausen (ed.), *Socialization and Society*, Little, Brown and Company, Boston, 1969.

Smith, Marshall S., "Discussion," *CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten*, proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973a.

-----, "Large Scale 'Experimentation' in Education," in A. Rivlin and T. M. Timpane (eds.), *Planned Variation in Education: Should We Give Up or Try Harder*? Brookings Institution, Washington, D.C. (forthcoming).

-----, *Some Short-Term Effects of Project Head Start: A Preliminary Report on the Second Year of Planned Variation; 1970-1971*, Huron Institute, Cambridge, Massachusetts, 1973b.

-----, and Joan S. Bissell, "Report Analysis: The Impact of Head Start," *Harvard Educational Review*, Vol. 40, No. 1, winter 1970, pp. 51-104.

Snedecor, George W., and William G. Cochran, *Statistical Methods*, 6th ed., The Iowa State University Press, Ames, 1967.

Soar, R. S., *Follow-Through Classroom Process Measurement and Pupil Growth (1970-71)*, Final Report, University of Florida, Gainesville, June 1973.

-----, and Ruth M. Soar, *Research Reports*, "Classroom Behavior, Pupil Characteristics and Pupil Growth for the School Year and the Summer," Institute for Development of Human Resources, University of Florida, Gainesville, December 1973.

Sokolow, J., and E. J. Taylor, "Report of a National Field Trial of a Method for Functional Disability Evaluation," *Journal of Chronic Diseases*, November-December 1967, pp. 897-909.

Spivak, G., and M. Shure, *Social Adjustment of Young Children*, Jossey-Bass, San Francisco, 1973.

Stallings et al., *Follow-Through Program Classroom Observation Evaluation 1971-72*, Stanford Research Institute, Menlo Park, California. Prepared for Bureau of Elementary and Secondary Education, U.S. Office of Education, Department of Health, Education, and Welfare, August 1973.

Stanford Research Institute, *Classroom Observation Study of Implementation in Head Start Planned Variation, 1970-1971*, Final Report, Menlo Park, California, August 10, 1973.

Stearns, Marion, "Planned Variation: What is the Question?" paper presented at Brookings Conference on Planned Variation in Education: Should We Give Up or Try Harder? Washington, D.C., April 1973.

-----, *Report on Preschool Programs: The Effects of Preschool Programs on Disadvantaged Children and Their Families*, Final Report, Office of Child Development, Department of Health, Education, and Welfare, Washington, D.C., 1974.

Sullivan, D. F., *Conceptual Problems in Developing an Index of Health*, National Center for Health Statistics, Publication No. 1000, Series 2, Number 17, Public Health Service, Washington, D.C., May 1966.

Sulzer, J., "Significance of Iron Deficiencies: Effects of Iron Deficiency on Psychological Tests in Children," in *Extent and Meanings of Iron Deficiency in the U.S.: Summary Proceedings of a Workshop*, Food and Nutrition Board, National Academy of Sciences, Washington, D.C., March 8-9, 1971.

System Development Corporation, "Effects of Different Head Start Program Approaches on Children of Different Characteristics: Report on Analysis of Data from 1968-1969 National Evaluation," Santa Monica, California, Prepared for Project Head Start, Office of Child Development, U.S. Department of Health, Education, and Welfare, May 1972a.

-----, "Effects of Different Head Start Program Approaches on Children of Different Characteristics: Report on Analysis of Data from 1966-1967 and 1967-1968 National Evaluations," Santa Monica, August 1972b.

Tanaka, M., and C. Massad, "Language Comprehension and Performance," *CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten*, proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973.

Tang, P. C., "The Power Function of the Analysis of Variance Tests with Tables and Illustrations of Their Use," *Statistics Research Memorandum*, Vol. 2, 1938, pp. 126-149.

Tatsuoka, M. M., *Multivariate Analysis; Techniques for Educational and Psychological Research*, John Wiley, New York, 1971.

Terrel, G., Jr., K. Durkin, and M. Wesley, "Social Class and the Nature of the Incentive in Discrimination Learning," *Journal of Abnormal and Social Psychology*, Vol. 59, 1959, pp. 270-272.

Terrel, G., Jr., and W. Kennedy, "Discrimination Learning and Transposition in Children as a Function of the Nature of the Reward," *Journal of Abnormal and Social Psychology*, Vol. 53, 1957, pp. 257-260.

Thorndike, Robert L., "Regression Fallacies in the Matched Groups Experiment," *Psychometrika*, Vol. 7, No. 2, June 1942, pp. 85-102.

Tiku, M. L., "Tables of the Power of the F-Test," *Journal of the American Statistical Association*, Vol. 62, 1967, pp. 525-539.

Tizard, Barbara et al., "Environmental Effects on Language Development: A Study of Young Children in Long-Stay Residential Nurseries," *Child Development*, Vol. 43, 1972, pp. 337-358.

Tukey, John W., *Exploratory Data Analysis*, Vols. 1-3, Addison-Wesley Publishing Company, Reading, Massachusetts, 1970.

U.S. Bureau of Census, *1970 Census Users Guide*, Washington, D.C., 1970.

U.S. Commission on Civil Rights, "Teachers and Students, Report V: Mexican American Education Study," Washington, D.C., 1973.

Vernon, M. D., "The Functions of Schemata in Perceiving," *Psychological Review*, Vol. 62, 1955, pp. 180-192.

Walker, Deborah Klein, *Socioemotional Measures for Preschool and Kindergarten Children*, Jossey-Bass, San Francisco, 1973.

- [illegible] - et al., *The Quality of the Head Start Planned Variations Data*, 2 [illegible]., Huron Institute, Cambridge, Massachusetts, 1973.

Ward, W. C., "Disadvantaged Children and Their First School Experiences: Development of Self-Regulatory Behaviors," Educational Testing Service, Head Start Longitudinal Study, 1973.

Watts, H. W., and D. L. Horner, *The Educational Benefits of Head Start: A Quantitative Analysis*, University of Wisconsin, Institute for Research on Poverty, Madison, 1968.

Weiner, B., "Attribution Theory, Achievement Motivation, and the Educational Process," *Review of Educational Research*, Vol. 42, 1972, pp. 203-215.

Weinstein, E. A., "The Development of Interpersonal Competence," in D. Goslin (ed.), *Handbook of Socialization Theory and Research*, Rand McNally and Company, Chicago, 1969.

Weisberg, Herbert I., *Short-Term Cognitive Effects of Head Start Programs: A Report on the Third Year of Planned Variation--1971-1972*, Huron Institute, Cambridge, Massachusetts, August 1973.

Werts, Charles E., and Robert L. Linn, "A General Linear Model for Studying Growth," *Psychological Bulletin*, Vol. 73, No. 1, 1970, pp. 17-22.

White, Sheldon et al., *Federal Programs for Young Children*, prepared for the Department of Health, Education, and Welfare, Contract No. OS-71-170, Huron Institute, Cambridge, Massachusetts, 1973.

Williams, F., "Language, Attitude, and Social Change," in F. Williams (ed.) *Language and Poverty*, Markham Press, Chicago, 1970a.

-----, "Some Preliminaries and Prospects," in F. Williams (ed.), *Language and Poverty*, Markham Press, Chicago 1970b.

-----, and R. Naremore, "On the Functional Analysis of Social Class Differences in Modes of Speech," *Speech Monographs*, Vol. 36, 1969, pp. 77-108.

Winer, B. J., *Statistical Principles in Experimental Design* (2d. ed.), McGraw-Hill Book Company, New York, 1971.

Wolff, M., and A. Stein, "Head Start Six Months Later," in J. L. Frost (ed.), *Early Childhood Education Rediscovered: Readings*, Holt, Rinehart and Winston, New York, 1967.

World Health Organization, *Measurements of Levels of Health: Report of a Study Group*, WHO Technical Report Series No. 137, Geneva, 1957.

Wright, H. F., "Observation Child Study," in Paul Mussen (ed.), *Handbook of Research Methods in Child Development*, John Wiley and Son, Inc., New York, 1960, pp. 71-139.

-----, and R. G. Barker, *Methods in Psychological Ecology*, University of Kansas, Lawrence, 1950.

Yee, A. H., "Interpersonal Attitudes of Teachers and Advantaged and Disadvantaged Pupils," *Journal of Human Resources*, Vol. 3, No. 3, 1968, pp. 327-345.

Young, C., and F. McConnell, "Retardation of Vocabulatory Development in Hard of Hearing Children," *Exceptional Child*, 1957, pp. 368-370.

Zigler, E., "The Environmental Mystique: Training the Intellect Versus Development of the Child," *Childhood Education*, 1970, pp. 402-412.

-----, "Myths and Facts: A Guide for Policymakers," *Compact*, July/August 1973a, pp. 18-21.

-----, "Project Head Start: Success or Failure," *Learning*, Vol. 1, No. 7, 1973b, pp. 43-47.

-----, and E. Butterfield, "Motivational Aspects of Changes in IQ Test Performance of Culturally Deprived Nursery School Children," *Child Development*, Vol. 39, 1968, pp. 1-14.

Zigler, E., and J. deLabry, "Concept-Switching in Middle-Class, Lower-Class and Retarded Children," *Journal of Abnormal and Social Psychology*, Vol. 65, 1962, pp. 267-273.

Ziller, R. C., "A Helical Theory of Personal Change," *Journal for the Theory of Social Behavior*, Vol. 1, 1971, pp. 33-74.

----- et al., "Self-Esteem: A Self-Social Construct," *Journal of Consulting Psychology*, Vol. 33, 1969, pp. 84-95.